

Department of Cell and Molecular Biology  
Karolinska Institutet, Stockholm, Sweden

# **Genome and transcriptome studies of the protozoan parasites *Trypanosoma cruzi* and *Giardia intestinalis***

Oscar Franzén



**Karolinska  
Institutet**

Stockholm 2012

*Cover:* (Front) Artwork showing replicating *Giardia intestinalis* trophozoites and *Trypanosoma cruzi* epimastigotes. DAPI staining (blue) shows the two nuclei of the former parasite. Original micrographs: J. Jerlström-Hultqvist and M. Ferella.

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet. Printed by Larserics Digital Print AB.

© Oscar Franzén, 2012

ISBN 978-91-7457-860-7

*To my parents...*



## Abstract

*Trypanosoma cruzi* and *Giardia intestinalis* are two human pathogens and protozoan parasites responsible for the diseases Chagas disease and giardiasis, respectively. Both diseases cause suffering and illness in several million individuals. The former disease occurs primarily in South America and Central America, and the latter disease occurs worldwide. Current therapeutics are toxic and lack efficacy, and potential vaccines are far from the market. Increased knowledge about the biology of these parasites is essential for drug and vaccine development, and new diagnostic tests. In this thesis, high-throughput sequencing was applied together with extensive bioinformatic analyses to yield insights into the biology and evolution of *Trypanosoma cruzi* and *Giardia intestinalis*. Bioinformatics analysis of DNA and RNA sequences was performed to identify features that may be of importance for parasite biology and functional characterization. This thesis is based on five papers (*i-v*). Paper *i* and *ii* describe comparative genome studies of three distinct genotypes of *Giardia intestinalis* (A, B and E). The genome-wide divergence between A and B was 23% and 13% between A and E. 4557 groups of three-way orthologs were defined across the three genomes. 5 to 38 genotype-specific genes were identified, along with genomic rearrangements. Genes encoding surface antigens, *vsps*, had undergone extensive diversification in the three genotypes. Several bacterial gene transfers were identified, one of which encoded an acetyltransferase protein in the E genotype. Paper *iii* describes a genome comparison of the human infecting *Trypanosoma cruzi* with the bat-restricted subspecies *Trypanosoma cruzi marinkellei*. The human infecting parasite had an 11% larger genome, and was found to have expanded repertoires of sequences related to surface antigens. The two parasites had a shared ‘core’ gene complement. One recent horizontal gene transfer was identified in *T. c. marinkellei*, representing a eukaryote-to-eukaryote transfer from a photosynthesizing organism. Paper *iv* describes the repertoire of small non-coding RNAs in *Trypanosoma cruzi* epimastigotes. Sequenced small RNAs were in the size range 16 to 61 nucleotides, and the majority were derived from transfer RNAs and other non-coding RNAs. 92 novel transcribed loci were identified in the genome, 79 of which were without similarity to known RNA classes. One population of small RNAs were derived from protein-coding genes. Paper *v* describes transcriptome analysis using paired-end RNA-Seq of three distinct genotypes of *Giardia intestinalis* (A, B and E). Gene expression profiles recapitulated the known phylogeny of the examined genotypes, and 61 to 176 genes were differentially expressed. 49,027 distinct polyadenylation sites were mapped and compared, and the median 3’UTR length was 80 nucleotides (A). One 36-nt novel intron was identified and the previously reported introns (5) were confirmed.

# List of Publications

This thesis is based on the following articles and manuscripts that will be referred to in the text by their roman numerals.

- i **Franzén O**, Jerlström-Hultqvist J, Castro E, Sherwood E, Ankarklev J, Reiner DS, Palm D, Andersson JO, Andersson B, Svärd SG. **Draft Genome Sequencing of *Giardia intestinalis* Assemblage B Isolate GS: Is Human Giardiasis Caused by Two Different Species?**. *PLoS Pathogens*, 2009, Aug;5(8):e1000560.
- ii Jerlström-Hultqvist J, **Franzén O**, Ankarklev J, Xu F, Nohýnková E, Andersson JO, Svärd SG, Andersson B. **Genome analysis and comparative genomics of a *Giardia intestinalis* assemblage E isolate**. *BMC Genomics*, 2010, 11:543.
- iii **Franzén O**, Talavera-López C, Ochaya S, Butler CE, Messenger LA, Lewis MD, Llewellyn MS, Marinkelle CJ, Tyler KM, Miles MA, Andersson B. **Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages with the bat-restricted subspecies *T. cruzi marinkellei***. *BMC Genomics*. Accepted (in press).
- iv **Franzén O** <sup>†</sup>, Arner E <sup>†</sup>, Ferella M, Nilsson D, Respuela P, Carninci P, Hayashizaki Y, Åslund L, Andersson B, Daub CO. **The Short Non-Coding Transcriptome of the Protozoan Parasite *Trypanosoma cruzi***. *PLoS Neglected Tropical Diseases*, 2011, 5(8): e1283.  
<sup>†</sup> Equal contribution.
- v **Franzén O** <sup>‡</sup>, Jerlström-Hultqvist J <sup>‡</sup>, Einarsson E, Ankarklev J, Ferella M, Andersson B, Svärd SG. **Transcriptome Profiling of *Giardia intestinalis* Using Strand-specific RNA-Seq**. *Submitted manuscript*.  
<sup>‡</sup> Equal contribution.

## Other Publications

During the course of my doctoral studies, I also performed or participated in the following studies.

- **Franzén O**, Ochaya O, Sherwood E, Lewis MD, Llewellyn MS, Miles MA, Andersson B. **Shotgun Sequencing Analysis of *Trypanosoma cruzi* I Sylvio X10/1 and Comparison with *T. cruzi* VI CL Brener.** *PLoS Neglected Tropical Diseases*, 2010, 5(3): e984.
- Jiao W, Masich S, **Franzén O**, Shupliakov O. **Two pools of vesicles associated with the presynaptic cytosolic projection in *Drosophila* neuromuscular junctions.** *Journal of Structural Biology*, 2010, 389–394.
- Messenger LA, Llewellyn MS, Bhattacharyya T, **Franzén O**, Lewis MD, Ramírez JD, Carrasco HJ, Andersson B, Miles MA. **Multiple Mitochondrial Introgression Events and Heteroplasmy in *Trypanosoma cruzi* Revealed by Maxicircle MLST and Next Generation Sequencing.** *PLoS Neglected Tropical Diseases*, 2012, 6(4): e1584.





# Contents

<b>1</b>	<b>Synopsis</b>	<b>1</b>
<b>2</b>	<b>Introduction</b>	<b>3</b>
2.1	Current State of Genome Sequencing . . . . .	3
2.2	RNA-Seq – A Method to Read the Transcriptome at Single Nucleotide Resolution . . . . .	5
2.3	<i>Giardia intestinalis</i> – A Gastrointestinal Parasite of Humans and Animals . . . . .	7
2.3.1	Cell Biology and Life Cycle: Regression and Simplicity .	7
2.3.2	Is <i>G. intestinalis</i> a Primitive Eukaryote or Highly Adapted Towards Parasitism? . . . . .	9
2.3.3	Intraspecific Taxonomy: Two Genotypes Infect Humans	10
2.3.4	The Streamlined Genome of <i>G. intestinalis</i> Reveals Many Parasite-specific Genes . . . . .	11
2.3.5	Unexpected Low Heterozygosity in a Polyploid Organism	13
2.3.6	Promiscuous Transcription due to Loose Transcriptional Regulation? . . . . .	13
2.4	<i>Trypanosoma cruzi</i> – A Pathogen Transmitted by Blood-sucking Insects and Cause of Systemic Illness . . . . .	15
2.4.1	<i>T. cruzi</i> is Transmitted by a Wide Range of Insect Vectors	16
2.4.2	The Complex Life Cycle of <i>T. cruzi</i> . . . . .	17
2.4.3	The Population Structure of <i>T. cruzi</i> is Wide, Complex and Contains Signatures of Ancestral Hybridization . .	18
2.4.4	Was the Ancestor of <i>T. cruzi sensu stricto</i> a Bat Trypanosome? . . . . .	19
2.4.5	An Unusual Amount of Genomic Redundancy . . . . .	21
2.4.6	Lack of RNA interference in <i>T. cruzi</i> but not in <i>T. brucei</i>	23
<b>3</b>	<b>Aims</b>	<b>25</b>
3.1	Specific aims . . . . .	25
<b>4</b>	<b>Present investigation</b>	<b>26</b>
4.1	Paper <i>i</i> and <i>ii</i> : Genome Comparison of Three Distinct Isolates of <i>Giardia intestinalis</i> . . . . .	26
4.1.1	The Core Genome and Isolate-specific Genes . . . . .	28
4.1.2	Structural Variation and Overall Divergence . . . . .	29

4.2	Paper <i>iii</i> : Genome Comparison of <i>Trypanosoma cruzi sensu stricto</i> With the Bat-restricted Subspecies <i>T. cruzi marinkellei</i>	30
4.2.1	Evidence of a Recent Eukaryote-to-Eukaryote Horizontal Gene Transfer . . . . .	32
4.2.2	<i>Trypanosoma cruzi marinkellei</i> has Capacity to Invade non-Bat Epithelial Cells . . . . .	32
4.3	Paper <i>iv</i> : The Small RNA Component of the <i>Trypanosoma cruzi</i> Transcriptome . . . . .	34
4.4	Paper <i>v</i> : Transcriptome Profiling at Single Nucleotide Resolution of Diverged <i>G. intestinalis</i> Genotypes Using Paired-end RNA-Seq . . . . .	36
4.4.1	Low Biological and Technical Variation . . . . .	37
4.4.2	Gene Expression Levels Recapitulate the Known Phylogeny . . . . .	37
4.4.3	Promiscuous Polyadenylation Sites and Unusual <i>cis</i> -acting Signals . . . . .	39
4.4.4	Biallelic Transcription and Correlation of Allele Expression and Allele Dosage . . . . .	41
4.4.5	Only Six Genes are <i>cis</i> -spliced . . . . .	41
<b>5</b>	<b>Concluding remarks</b>	<b>44</b>
5.0.6	Comparative Genomics . . . . .	44
5.0.7	The Short Transcriptome of <i>T. cruzi</i> . . . . .	45
5.0.8	Comparative Transcriptomics of <i>G. intestinalis</i> . . . . .	45
<b>6</b>	<b>Populärvetenskaplig sammanfattning</b>	<b>47</b>
<b>7</b>	<b>Acknowledgments</b>	<b>48</b>
<b>8</b>	<b>References</b>	<b>50</b>

## List of abbreviations

<b>AER</b>	Allelic expression ratio
<b>AR</b>	Allele ratio
<b>ASE</b>	Allele-specific expression
<b>bp</b>	Base pairs
<b>cDNA</b>	Complementary DNA
<b>CDS</b>	Coding sequence
<b>contig</b>	Contiguous sequence
<b>DNA</b>	Deoxyribonucleic acid
<b>DTU</b>	Discrete typing unit
<b>FPKM</b>	Fragments per kilobase of transcript per million fragments mapped
<b>LBA</b>	Long branch attraction
<b>nt</b>	Nucleotides
<b>ORF</b>	Open reading frame
<b>PAC</b>	Polyadenylation site cluster
<b>PCR</b>	Polymerase chain reaction
<b>polyA</b>	Polyadenylation
<b>PTU</b>	Polycistronic transcription unit
<b>RNA</b>	Ribonucleic acid
<b>RNA-Seq</b>	RNA-Sequencing
<b>RNAi</b>	RNA interference
<b>s.s.</b>	sensu stricto
<b>SAGE</b>	Serial analysis of gene expression
<b>SGS</b>	Second generation sequencing
<b>SNP</b>	Single nucleotide polymorphism
<b>sRNA</b>	Small RNA
<b>tsRNA</b>	tRNA-derived small RNA
<b>UTR</b>	Untranslated region



# 1 Synopsis

*Nothing in biology makes sense except in the light of evolution*

– T.G. Dobzhansky (1973), geneticist and evolutionary biologist.

Infectious diseases are leading causes of suffering and death of humans around the world, and have significant impact on daily lives of many million people. Parasites cause some of the worst and most neglected diseases, including Malaria, African- and American trypanosomiasis, Schistosomiasis and several others. These diseases are most prevalent in tropical regions of the world, and are often associated with poverty as well as being intrinsically “poverty promoting.” Unsafe drinking water, compromised hygiene and sanitary conditions or substandard housing are factors that facilitate disease. Many of the afflicted individuals have very limited access to health care. Several factors can be attributed to the lack of treatment options: neglected diseases attract little attention from pharmaceutical companies and first-world governments, often because companies are unable to regain investments in expensive basic research, drug development and clinical trials. Moreover, many parasites are difficult to study in the laboratory due to complex life cycles or because the parasites do not readily grow *in vitro*. Most neglected diseases do not cause acute outbreaks, and instead progress during many years and in the meantime cause debilitating illness and suffering.

In addition to the mission of improving human health, parasites often have unique or specialized biological features, which makes them excellent models for the study of eukaryotic evolution. Parasites provide a window into the biological and social evolution of our own species, since many parasites have co-evolved together with the *Homo* lineage for many millions of years; this appears to be the situation for the worms *Trichuris trichiura* and *Enterobius* [1, 2]; other parasites such as *Trypanosoma cruzi* tell a shorter story of human co-evolution. Parasite-host co-evolution has likely resulted in reciprocal adaptations with complex evolutionary consequences, for example the favouring of specific genetic processes such as recombination that operate to create new genotypes to which the host is not adapted. Another example of co-evolution can be found in our own species, where a trypanolytic factor encoded by our genome is likely an ancestral adaptation against trypanosomatid parasites of

the *T. brucei* clade [3].

Second generation sequencing enables cost-efficient and rapid acquisition of large data sets covering diverse biological aspects, including but not limited to genome, metagenome, epigenome and transcriptome studies. An important target of the new technologies is human parasites, aiming to deepen our understanding of the underlying biology of these pathogens, and their evolutionary trajectories. Such efforts may reveal signatures relating to how parasitism evolved. Second generation sequencing has already facilitated key insights into the molecular organization of these organisms, and rapidly enables advancement of functional studies, identification of drug targets and formation of new hypotheses. Old questions relating to pathogenicity, epidemiology and genetics can be addressed with the new tools and may ultimately lead to insights that pave the way for better treatment strategies for neglected and tropical diseases.



## 2 Introduction

*Fossil bones and footsteps and ruined homes are the solid facts of history, but the surest hints, the most enduring signs, lie in those miniscule genes. For a moment we protect them with our lives, then like relay runners with a baton, we pass them on to be carried by our descendants. There is a poetry in genetics which is more difficult to discern in broken bones, and genes are the only unbroken living thread that weaves back and forth through all those boneyards.*

– J. Kingdon, biologist and science author (1996).

### 2.1 Current State of Genome Sequencing

Genomes vary enormously in size, with many of the size differences being caused by repeated sequences (Figure 1). Deoxyribonucleic acid (DNA) sequencing techniques are limited to reading short pieces of DNA. The most common strategy to overcome this limitation is shotgun sequencing. The technique involves random fragmentation of chromosomes to a redundant mix of small DNA fragments. These fragments can subsequently be sequenced, resulting in a mix of sequences of forward and reverse directions, representing the original chromosomes. By using the overlap of these short sequences, computer programs can reconstruct millions of short sequences into longer sequences (contigs) – a process referred to as genome assembly. While simple in theory, the task becomes less straightforward due to the following obstacles: *(i)* the large size of most genomes, especially those of eukaryotes; *(ii)* the vast amount of sequence data needed to achieve sufficient redundancy, i.e. the genome “coverage”; *(iii)* the fact that many DNA fragments are identical or close to identical (repeats); *(iv)* heterozygosity, i.e. single nucleotide polymorphisms between homologous chromosomes; and *(v)* sequence errors. Apart from these issues, genomes can exhibit aneuploidy and complex karyotypes – all of which make the assembly task more difficult. Sanger sequencing is referred to as ‘first generation sequencing,’ and has been used to sequence large and complex genomes, e.g. the human genome. Sanger sequencing can

read up to 1000 base pairs (bp) of a DNA fragment. Despite being relatively old, it is still the most common technique for low-throughput applications, e.g. DNA amplified from Polymerase Chain Reaction (PCR). A comparison of the repeat content in relation to assembly consistency of various draft or complete genomes is shown in Figure 2.

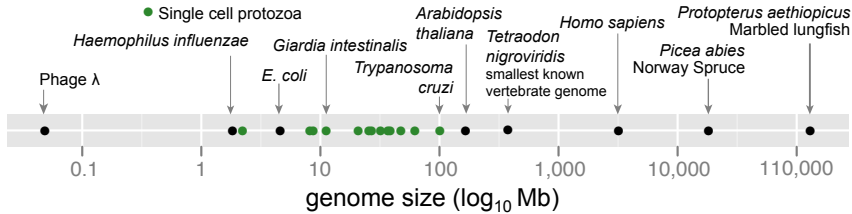


Figure 1: Genome sizes on a logarithmic scale. Phage λ has a genome of 48,502 bp and it was determined by Sanger *et al.* in 1982 [4]. Most genomes of sequenced, parasitic protozoa are between 4 to 100 Mb in size. The genome size of *T. cruzi* refers to both haplotypes of the CL Brener strain. For the other eukaryotes the genome size refers to the haploid state. (Green dots) Single cell protozoa. Genomes of mammals are at least an order of magnitude larger. Marbled lungfish has the largest genome of any known organism (130,000 Mb) [5].

Second generation sequencing (SGS) offers higher throughput, at lower cost per base, but yields shorter sequences (reads). Short reads are often a problem for determining a genome sequence, as most genomes contain repetitive sequences longer than the read length [6]. Paired-end protocols have been developed to tackle repeats, and allow sequencing of longer DNA fragments from both ends in order to bridge repeats. Paired-end reads of various sizes can subsequently be used to link contigs into scaffolds. Hence, ‘scaffolds’ is the genomics term for contigs that are ordered and oriented. Several SGS techniques or platforms are available, for example Roche/454 sequencing, developed from pyrosequencing. The platform from Illumina provides significantly higher throughput, albeit at shorter read lengths. Most genome sequencing efforts combine data from different platforms to overcome their respective limitations. Repetitive sequences are currently the major bottleneck in genome sequencing projects, especially since most eukaryotes contain various classes of repeats, e.g. retroelements and segmental duplications. Future developments are anticipated to improve genome sequences, including Pacific Biosciences and perhaps more distant, Nanopore sequencing [7].



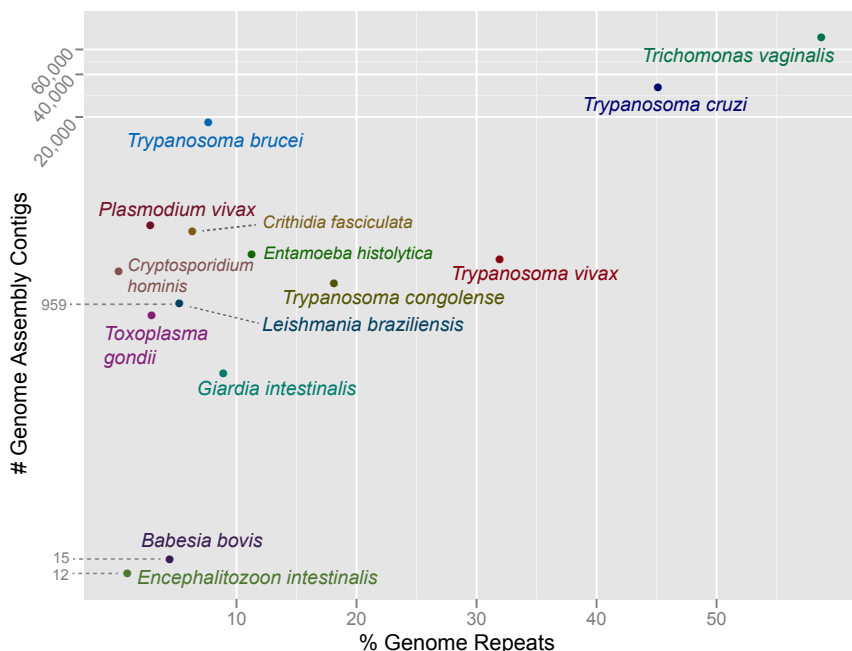


Figure 2: Comparison of genome repeat-content and assembly fragmentation of various eukaryotic pathogens from EupathDB [8]. Repeat libraries were established for each genome using RepeatScout [9] (repeats >500 bp of length), and contigs >200 bp of each genome were searched using RepeatMasker [10]. (X-axis) Percentage of the genome (sum of contig lengths) present in repeats. (Y-axis) Contig count of the assembly (logarithmic scale). Strains: *B. bovis* (T2Bo); *C. fasciculata* (Cf-C1); *C. hominis* (TU502); *E. intestinalis* (ATCC 50506); *E. histolytica* (HM-1:IMSS); *G. intestinalis* (WB); *L. braziliensis* (M2903); *P. vivax* (SaI-1); *T. gondii* (GT1); *T. vaginalis* (G3); *T. brucei* (427); *T. congolense* (IL3000); *T. cruzi* (CL Brener); and *T. vivax* (Y486).

## 2.2 RNA-Seq – A Method to Read the Transcriptome at Single Nucleotide Resolution

The field of transcriptomics aims to describe all transcripts in a cell or tissue and to determine the features of these transcripts: for example 5' and 3' end structures, splicing patterns and quantitative information about transcript levels. Transcriptome sequencing (RNA-Seq) relies on deep sequencing of fragmented cDNA libraries, achieving its quantitative properties by the amounts of sequence reads of a particular transcript [11], i.e. abundant transcripts yield more reads whereas more rare transcripts yield fewer. Therefore,

the final digital expression values are based on simple counting statistics, which are often normalized by gene length and the number of sequences generated by the instrument. While microarray data often require complicated normalization procedures, processing of RNA-Seq data is relatively simple and straightforward. RNA-Seq results in lower background noise than microarrays [12]. Any high-throughput sequencing platform can be utilized, but Illumina has been the most widely adopted and is currently the best supported in terms of bioinformatics software. RNA-Seq enables *ab initio* discovery of new and rare transcripts and splicing patterns, which cannot be observed on standard microarrays. Several software programs are freely available to process RNA-Seq data. These programs rely on optimized algorithms to map (align) large quantities of sequence data, and at the same time consider polymorphisms and sequence errors. RNA-Seq has rapidly become widely adopted and applied to various research questions, including differential expression analysis [13], small ribonucleic acid (RNA) discovery, allele-specific expression [14] as well as mapping 5' and 3' ends of genes [15].

Recent developments allow the generation of paired-end libraries, read lengths up to 150 nucleotides and strand specificity. These improvements allow going beyond the mRNA component of the transcriptome and sampling hidden transcriptional layers. Drawbacks of the RNA-Seq method include: (i) the cost of library preparation and sequencing; (ii) lack of user-friendly analysis pipelines and interfaces; (iii) RNA or cDNA must be fragmented into smaller pieces, usually between 100 to 500 nt; (iv) library preparation and fragment amplification may introduce artifacts or biases; (v) transcript coverage bias is common, i.e. coverage fluctuations along the 5' to 3' axis of the mRNA; (vi) long-time storage of RNA-Seq data sets is becoming increasingly difficult because of large data volumes; and (vii) certain downstream analysis tasks, e.g. discovery of rare transcript isoforms and splice variants, still suffer from many false positives due to artifactual chimeras or amplification biases from library preparation or sequencing. Future developments in single molecule sequencing may ameliorate such problems. Despite these limitations, RNA-Seq is likely to improve and provide novel insight into the transcriptomes of protozoans and other eukaryotes.



## 2.3 *Giardia intestinalis* – A Gastrointestinal Parasite of Humans and Animals

*Giardia intestinalis* is a protozoan, amitochondrial parasite and member of the diplomonad group of species, which includes other anaerobic or microaerophilic protozoans. The diplomonad group is part of the supergroup Excavata [16]. In the literature, the species names *G. intestinalis*, *G. duodenalis* and *G. lamblia* are used interchangeably and refer to the same organism. The parasite was discovered already in 1681 by the Dutch microscopist van Leeuwenhoek [17], and described in more detail in 1859 by the Czech physician Lambl [18]. *G. intestinalis* infects humans and animals and is one of the most prevalent gastrointestinal parasites worldwide [19]. *G. intestinalis* is a potential zoonotic pathogen since it can infect a broad range of mammals in addition to humans. In man, the parasite colonizes the upper part of the small intestine and adheres to the mucosa along the sides of villi. The infection causes diarrhea, and may lead to malnutrition and failure of children to thrive [20]. The disease is particularly a burden in developing countries, where compromised hygiene may increase transmission and cause endemic outbreaks. Local outbreaks do occasionally occur in developed countries, for example via the public water supply [21] or in day care centers [22]. In 2004 a large outbreak of *G. intestinalis* occurred in Bergen Norway, with altogether 1,300 laboratory-confirmed cases [23]. Since the outbreak certain individuals have had prolonged and recurring symptoms of giardiasis, with a profound impact on the quality of life [24]. Recent data indicate a putative relationship between irritable bowel syndrome and previous *G. intestinalis* infection [25].

As of 2004, *G. intestinalis* has been included in the WHO Neglected Disease Initiative [26].

### 2.3.1 Cell Biology and Life Cycle: Regression and Simplicity

In contrast to other protozoan parasites, *G. intestinalis* has a relatively simple life cycle, consisting of the dormant cyst stage and the replicative trophozoite stage. Trophozoites have a characteristic half pear-shaped morphology, and are 12-15  $\mu\text{m}$  long, 5-9  $\mu\text{m}$  wide and 1-2  $\mu\text{m}$  thick (Figure 3). Trophozoites have four pairs of flagella, which are anchored to the cytoskeleton. The parasite rotates around its longitudinal axis to create a forward propulsion force. The rotation causes the parasite to move at a speed of 12-40  $\mu\text{m/s}$  [27]. An adhesion disk is present on the ventral surface of the parasite, and

is used for anchoring to substrates. Hernández-Sánchez *et al.* reported that adhesion-deficient *G. intestinalis* had reduced capacity to establish infection in Mongolian gerbils [28]. When the parasite anchors to substrates, whether artificial or natural, the pattern of motion changes to more stable planar swimming [27]. Unusually compared to most eukaryotes, trophozoite cells contain two transcriptionally active nuclei [20]. Each nucleus contains a diploid to tetraploid set of the genome [29]. The biological significance of the polyploid genome is not clear, but it is a shared feature among many diplomonads (order *Diplomonadida*) and likely relates to the evolutionary history of the order. *G. intestinalis* has a well-defined endoplasmatic reticulum, which can form excretory vesicles [30]. *G. intestinalis* lacks a canonical Golgi apparatus and mitochondria. A vestigial organelle called mitosome is present. Mitosomes are double-membraned structures that appear to be involved in iron metabolism [31].

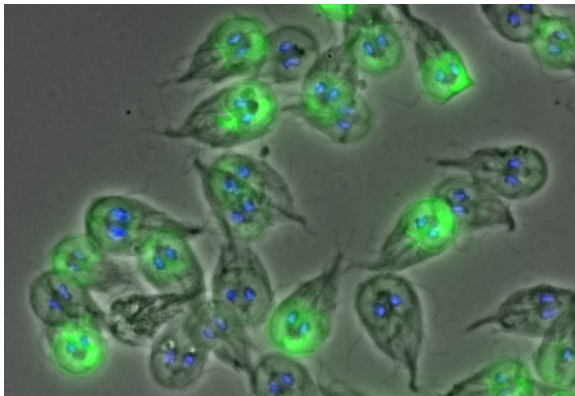


Figure 3: *G. intestinalis* trophozoites seen through a microscope. (Green) Tagged median body protein. (Blue) DAPI stained nuclei. Image credit: J. Jerlström-Hultqvist.

Cysts are the non-motile and metabolically dormant stage of the life cycle, and represent the infectious agents of giardiasis. Cysts can persist in the environment for prolonged periods and remain infectious, being encapsulated in a thick cyst wall of carbohydrate and protein. The most common route of transmission is the fecal-oral route, via contaminated food or water. Infection can also occur via person-to-person contact due to poor hygiene. The infectious dose can be as low as 10 cysts, as shown in a Texas prison “volunteer” population in 1954 [32]. Ingested cysts undergo excystation inside the host,

triggered by stomach acids. Cysts then rupture in the small intestine. Giardiasis is characterized by watery diarrhea, gastric pain and weight loss, and often but not always resolves spontaneously. The pathophysiology of giardiasis is poorly understood, but likely involves dysfunction of the epithelial cell barrier of the intestine and disturbances of the electrolyte balance [33]. The infection triggers apoptosis of host epithelial cells, and causes shortening of the brush border villi, all of which may contribute to diarrhoea [20]. There is also evidence that proteolytic enzymes released by *G. intestinalis* are involved in the disease [34]. Encystation is the process where trophozoites are transformed back to cysts, and it is triggered by the intestinal environment (high levels of bile, low cholesterol and/or shift in pH) [20].

### 2.3.2 Is *G. intestinalis* a Primitive Eukaryote or Highly Adapted Towards Parasitism?

Phylogenies based on nucleotide and protein sequences have consistently identified *G. intestinalis* as a basal eukaryote [35, 36, 37, 38]. This view has been corroborated by the apparent lack of some intracellular compartments (e.g. mitochondria, Golgi and peroxisomes) and an overall simplified, bacterial-like metabolism [19].

Roger *et al.* reported the finding of the mitochondrial-like gene *cpn60* in the genome of *G. intestinalis* [39]. The same year Hashimoto *et al.* reported the finding of a nuclear-encoded valyl-tRNA synthetase gene [40], which is regarded to be of mitochondrial origin in eukaryotes. A mitochondria-derived organelle, the mitosome, was later discovered [31]. Together these data suggest that *G. intestinalis* diverged after the endosymbiosis of the mitochondrial ancestor, but subsequently lost this feature, possibly as an adaptation to the microaerophilic life in the intestine. The finding of nucleoli also points in the direction of a typical eukaryote [41]. The early-branching position of *G. intestinalis* in phylogenetic trees has been questioned as an artifact caused by long-branch attraction (LBA) [42]. The problem of LBA arises when comparing taxa with variable evolutionary rates, which may lead to the artifactual early emergence of these taxa [43]. The effect of LBA may be mitigated by inclusion of additional species. Analysis of small nucleolar RNAs from Archaea and various unicellular eukaryotes has suggested that *G. intestinalis* emerged later than *Trypanosoma* and *Euglena* [44]. Altogether the current data of this parasite indicate that it has undergone reductive evolution and is highly adapted towards parasitism.

### 2.3.3 Intraspecific Taxonomy: Two Genotypes Infect Humans

*G. intestinalis* propagates via binary fission, i.e. it is an asexual process. Whether *G. intestinalis* participates in rare sexual events has been subject of debate, but there is currently no direct evidence for a sexual or parasexual cycle. Tibayrenc *et al.* showed that the parasite meets the criteria for a predominantly clonal population structure [45]. Nevertheless, recent data have suggested the possibility of infrequent genetic exchange [46, 47]. Many human-infective protozoans have sexual cycles or infrequently participate in genetic exchange; including *Trypanosoma cruzi* [48], *Toxoplasma gondii* [49] and *Plasmodium falciparum* [50]. Predominant clonal propagation does not preclude the existence of rare sexual events, but several questions are unresolved or inconsistent with a conventional sexual organism.

*G. intestinalis* has been suggested to comprise a species complex, consisting of eight distinct but morphologically indistinguishable genotypes (assemblages; Figure 4). Of the eight recognized assemblages (A to H), only two (A and B) infect humans as well as various non-human primates, cattle and many other animals [51]. Population studies have revealed further substructure of assemblage A, which can be subgrouped into AI, AII and AIII. Variation in pathogenicity among strains has been documented [52, 53], indicating a putative relationship between genotype and symptomatology. However, attempts to associate genotype with disease outcome have often been conflicting, and there is currently no certain relationship. In contrast, assemblage B has no clear subgrouping [51]. Only assemblage B has been used in experimental human infections [54].

Assemblages C to H are not associated with human infections, display stronger host-specificity and are less studied due to the difficulty to cultivate them *in vitro*. Parasites from assemblages C and D have been identified in dogs, wolves, coyotes and cats; E parasites have been found in cattle, sheep, pigs, goats and water buffalo; F parasites are reported mainly in cats; G parasites are mainly in rodents [51]. H parasites were relatively recently discovered in marine vertebrates [55]. The phylogenetic topology of the assemblages suggests that the extant A, E and F lineages share a common ancestor. It is possible that animal domestication provided opportunities for parasites to cross species boundaries and thereby adapt to new hosts.

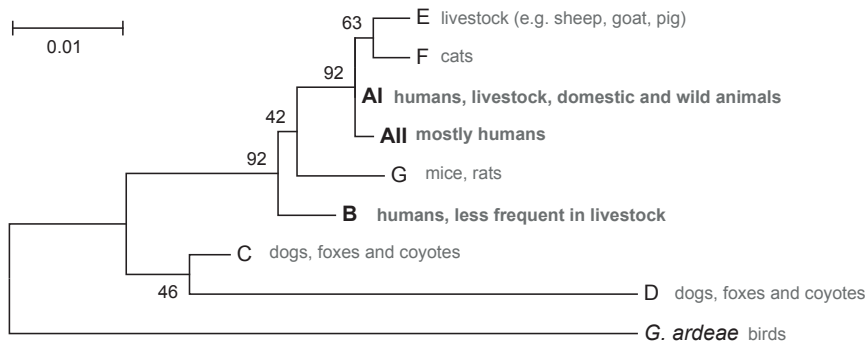


Figure 4: Maximum likelihood phylogeny of *G. intestinalis* assemblages A to G based on the *eflα* gene (nucleotide sequences). Sequences were aligned with ClustalW2 and the topology was inferred using the Tamura-Nei model as implemented in MEGA5 [56]. The scale bar refers to substitutions per site. Support values at branches were generated from 1000 bootstrap replicates. *G. ardeae*, a species found in birds [57], was used as an outgroup to support the phylogeny. Accession numbers of sequences used to infer the phylogeny; D14342.1, AF069573.1, AF069570.1, AF069574.1, AF069575.1, AF069571.1, AF069572.1, AF069568.1, AF069567.1.

Because of the different host species as well as genetic characteristics, reorganization of the assemblages into separate species has been proposed and is under debate [58]. The new species names for A and B are suggested to be *G. duodenalis* and *G. enterica*, respectively. Further studies and gathering of phenotypic data may shed light on the biology of non-human associated *Giardia* parasites.

### 2.3.4 The Streamlined Genome of *G. intestinalis* Reveals Many Parasite-specific Genes

One striking feature of *G. intestinalis* is the highly reduced genome, which is comprised of ~12 million base pairs (haploid size) distributed on five chromosomes [19]. Upcroft *et al.* reported a certain amount of karyotype variability in human and animal stocks [59], suggesting that the karyotype is not completely stable. Each nucleus of trophozoites contains a diploid to tetraploid set of the genome [60, 61, 29]. The genome of the assemblage A isolate WB was finished 2007 [62], and revealed sparse non-coding sequences, and with few exceptions intronless genes and simplified cellular components; including many bacterial- and archaeal-like enzymes. DNA synthesis, transcription, RNA processing and cell cycle components were found to be simple [62]. Since

the completion of the genome sequence, five genes containing introns have been found [62, 63, 64, 65]. The discovery of splicing in *G. intestinalis* was surprising and suggested that splicing was present at an early stage in the ancestral eukaryote. Recently, three *trans*-splicing events have been uncovered, i.e. the joining of physically distant exons into contiguous mature mRNAs [66, 65]. The documented events involved exons of the *dhc* and *hsp90* genes. The finding of such relatively complicated transcript maturation pathways further contradicts the view of *G. intestinalis* as a “fossil” eukaryote. The genome does encode proteins involved in meiosis [62], but these may have alternative functions.

Despite extensive efforts to annotate the *G. intestinalis* genome, ~58% of the genes are without a known function. These genes lack sequence similarity to other sequenced genomes and may represent *Giardia*-specific genes. The *G. intestinalis* genome encodes four multigene families: *nek* (encoding kinases), *p21.1* (encoding structural proteins, containing ankyrin motifs), *vsp* (surface antigens) and *hcmp* (putative surface antigens). Altogether these genes comprise 30% (3.6 Mbp/12 Mbp) of the genome [62], indicating that gene duplication has been a major evolutionary force. Most genes of these families display extensive heterogeneity, indicating significant divergence since the presumed gene duplication event. 198 *nek* genes are present in the genome, most of which are predicted to encode catalytically inert proteins [62]. Manning *et al.* found some *nek* proteins localized to distinct parts of the cytoskeleton and cytoplasm [67]. However, the precise roles of most of these proteins are unknown. Variant-specific surface antigens encoded by *vsp* genes cover the surface of the parasite and shield it from the host immune system [19]. Sequence analysis of *vsp* genes has suggested substructure, recombination and divergence among these genes [68]. Only one *vsp* is expressed at the cell surface at any given time [69], and switching occurs every 6-13 generations [70]. *vsp* switching occurs spontaneously, and is proposed to involve an epigenetic mechanism [71] and/or RNA interference [72].

The *G. intestinalis* genome contains three families of retrotransposons, of which two are potentially active and localized to telomeres and one is dead and present in interstitial genomic regions [73]. All three families of retrotransposons are long interspersed nuclear element (LINE)-like elements. Ullu *et al.* reported a population of small RNAs derived from the retrotransposon family GIL/Gen1 located in telomeres, and hypothesized that these may have a role in transposon-silencing [74].



### 2.3.5 Unexpected Low Heterozygosity in a Polyploid Organism

The WB isolate of *G. intestinalis* contained <0.01% heterozygosity as estimated from genomic reads [62]. Asexual organisms with diploid or higher genome ploidy would be expected to accumulate extensive genomic heterozygosity, i.e. single nucleotide polymorphisms between homologous nucleotide sites. The phenomenon is known as the Meselson effect and has been observed in bdelloid rotifers [75]. However, not all asexual organisms display high levels of heterozygosity. One prominent example is asexual lineages of *Daphnia*, which reduce heterozygosity via ameiotic recombination [76].

Poxleitner *et al.* used an episomal plasmid to demonstrate genetic exchange between nuclei of cysts; a process named diplomixis [77], and may partially explain how the parasite maintains low heterozygosity. Carpenter *et al.* showed that cyst formation occurs from a single trophozoite and not by fusion of two trophozoites [78]. The authors of the former study also concluded that nuclear sorting, i.e. each daughter cell receives a pair of identical nuclei, is not likely to be a mechanism by which *G. intestinalis* reduces heterozygosity. The precise mechanism likely involves gene conversion and/or homologous recombination and is yet-to-be described.

### 2.3.6 Promiscuous Transcription due to Loose Transcriptional Regulation?

*G. intestinalis* contains two nuclei with an equal amount of DNA, which has been shown by DAPI staining [60]. Uridine incorporation into RNA showed that both nuclei are transcriptionally active [60]. Recent data indicate that the two nuclei may not be completely identical: (i) Benchimol *et al.* showed that the two nuclei differ in nuclear pore number and distribution [79]; (ii) Tůmová *et al.* reported that the nuclei differ in both number and size of chromosomes [80]; (iii) a microRNA precursor was found in only one nucleus [81]; and (iv) Yang *et al.* reported allele-specific expression of one *vsp* [82].

Compared with other well-studied eukaryotes, the transcriptional apparatus is simple: 21 of 28 of the eukaryotic RNA polymerase polypeptides are present, but only 4 of the 12 general transcription initiation factors [83]. Sequencing of cDNA clones has found short 5' and 3' untranslated regions, sometimes only a few nucleotides [19]. Transcriptome profiling using Serial Analysis of Gene Expression (SAGE) and microarrays have uncovered a limited set of differentially expressed genes [84, 85]. Current data on transcription

in *G. intestinalis* indicate loose regulation at the transcriptional level. Only a few regulatory promoter elements have been discovered, mainly for developmental genes [86, 87]. Intergenic distances are short (the median is 103 bp) [62], leaving little space for regulatory elements. Analysis of promoters has failed to reveal shared motifs or regulatory elements, and suggested that promoters are degenerate [88, 89, 62]. An AT-rich sequence of 8 bp was found to be sufficient to drive transcription [90]. Hence, AT-richness appears to be the only prerequisite to initiate transcription, possibly explaining the abundance of pervasive transcription in this organism [91]. One consequence of this organization is bidirectional promoters, which contribute to pervasive transcription [91].

Drosha and Exportin-5 are two essential components of the microRNA-processing pathway, both of which are missing in *G. intestinalis* [92]. However, the parasite has Dicer and Argonaut homologs. *G. intestinalis* Dicer has been cloned and shown to produce RNA fragments between 25 to 27 nucleotides *in vitro* [93], although with lower affinity for its small RNA products compared with the human homolog. A recent study used antisense-ribozyme RNA in giardavirus-infected trophozoites to knockdown expression of the mRNA encoding the Argonaut protein [92]. The authors found that knockdown of Argonaut mRNA inhibited trophozoite replication, and concluded that Argonaut has an important role in the parasite. Moreover, the same study found a snoRNA-derived small RNA of 26-nt length produced by Dicer, and localized it to the cytoplasm. Target sites of these small RNAs were identified in *vsp* genes. An independent study similarly reported *vsp* regulation by RNA interference [72]. Altogether, the RNA interference (RNAi) apparatus seems to not be completely analogous with that found in metazoans, which is also supported by the fact that giardavirus (a double-stranded RNA virus) can replicate in certain strains of the parasite [94].

## 2.4 *Trypanosoma cruzi* – A Pathogen Transmitted by Blood-sucking Insects and Cause of Systemic Illness

*Trypanosoma cruzi* (*T. cruzi*) is a protozoan parasite and causative agent of Chagas disease (American trypanosomiasis), both discovered and described by Carlos Chagas in 1909 [95]. Chagas disease is a zoonosis, affecting ~8 million people mainly in rural and peri-urban areas of Mexico, Central America and South America [96]. A wide range of insect vectors facilitate transmission of *T. cruzi*, and the endemic range of both the parasite and its vectors stretches from southern United States to Argentinean Patagonia [97]. Chagas disease also occurs in non-endemic countries, because of migratory influx from endemic countries in Latin America [98]. More than 300,000 individuals are currently estimated to carry the infection in the United States and >80,000 in Europe [97]. However, since the natural vectors are not present, the disease is mainly confined to the infected individuals and accidental transmission via blood transfusion or organ transplant.

Chagas disease is a chronic and systemic illness [97]. The parasite has likely existed among animals in the Americas for millions of years, as concluded from its wide geographical distribution and host range [99]. Another line of evidence comes from observations of pathology, where domestic animals and humans often display pathology from the infection, in contrast to wild animals where pathology has not been recorded. This suggests that wild animals and the parasite co-evolved, which led to attenuated virulence. Recovered *T. cruzi* DNA from 9000-year-old mummies indicates that the disease has been troubling humans for extensive time [100]. In humans, the acute phase of Chagas disease is often asymptomatic and lasts from weeks to months. If symptoms do occur during this phase, they are benign (fever, swollen lymph glands and occasionally, local inflammatory reaction at the bite site) [97]. During the acute phase, *T. cruzi* can infect any nucleated cell of the host and parasites may be found in the blood. The immune system eventually reduces the parasitaemia, but does not clear the infection completely and the individual can remain asymptomatic for years or decades. At the chronic stage, parasites are still present in specific tissues, for example, muscle or enteric ganglia. Several years after entering the chronic stage, 20-30% of the individuals develop irreversible lesions of the heart, colon and/or oesophagus, and in some cases the peripheral nervous system [97]. The main lesions of Chagas disease are focal and extensive myocardial fibrosis, driven by a latent inflammatory response. Cardiomyopa-

thy is manifested by cardiac arrhythmias, apical aneurysm, congestive heart failure, thromboembolism and sudden cardiac arrest [97]. Chagasic pathology of colon and oesophagus is referred to as the digestive form of the disease, and is caused by destruction of enteric ganglia. The end result is segmental paralysis of the colon and/or oesophagus. Digestive Chagas appears to be more prevalent south of the Amazon basin (Argentina, Brazil, Bolivia and Chile) [99].

Treatment of Chagas disease is currently limited to two drugs introduced over 40 years ago, nifurtimox and benznidazole [97]. The drugs appear to have variable efficacy, require long treatment periods (60 to 90 days) and do not have an effect towards advanced Chagas disease. The drugs can give rise to severe side effects, including kidney and liver failure. Nifurtimox can also cause neurological disturbances and seizures. Despite many promising new drug targets and leads, for example posaconazole [101], few candidates have moved beyond the discovery phase – likely due to limited funding from companies and governments.

#### 2.4.1 *T. cruzi* is Transmitted by a Wide Range of Insect Vectors

The parasite, *T. cruzi*, belongs to the kinetoplastid group, which also includes the human parasites *Leishmania* spp. and *T. brucei*. *T. brucei* is indigenous to Africa and *Leishmania* spp. can be found worldwide. Kinetoplastid parasites exhibit some unusual molecular processes, such as RNA editing [102], *trans*-splicing [103] and antigenic variation [104]. *T. cruzi* is transmitted by several different species of insect vectors, mainly of the genera *Triatoma*, *Panstrongylus* and *Rhodnius* (Hemiptera; Reduviidae). The first entomological description of a Triatomine, *Triatoma rubrofasciata*, was performed already in 1773 by the scientist De Geer [105]. The most important vectors for human transmission are *Triatoma infestans*, *Rhodnius prolixus* and *Triatoma dimidiata* [97]. Vector species differ in regional distribution; for example, *T. infestans* has been the most important vector of sub-Amazonian regions, whereas *Rhodnius prolixus* is the predominant vector in northern Latin America. The insects are hematophagous bugs that feed on vertebrate blood, causing transmission of the parasite. Insects often live in cracks of poor quality rural homes or huts, and emerge at night, biting people near the eye or mouth. Many different mammalian hosts act as parasite reservoirs and thereby sustain the transmission cycle of *T. cruzi*. More than 150 species of wild (e.g. armadillos, opossums and raccoons) and domestic (e.g. dogs,

cats and guinea pigs) animals can act as *T. cruzi* reservoirs. The disease can also be transmitted via non-vectorial mechanisms, including blood transfusion [106] and organ transplant [107] as well as congenital transfer from mother to fetus [108, 109]. Oral transmission is possible via ingested food or liquid [110], and is generally associated with massive parasite proliferation, with severe and acute clinical manifestations and high rate of mortality [111]. However, oral outbreaks are rare but have been documented [112]. Even more rarely, individuals have become infected by accidents in the laboratory [113]. Regional differences in disease severity have often been suspected and may be due to parasite genotype, host genetics, transmission cycles and control programs [99].

Control measures for Chagas disease include chemical insect control, improvement of housing conditions and education. Blood can be screened before transfusion using serological tests. Most but not all Latin American countries have implemented mandatory serology tests for blood donors. Vector control programs involving spraying of insecticides on houses and buildings, have largely been successful. For example, the “Southern Cone Initiative” has reduced Chagas transmission rates in the South Cone of the continent by disrupting transmission via the vector *Triatomina infestans* [114].

It is possible that Charles Darwin (1809-1882) contracted Chagas disease during his journey to the Americas, as suggested from descriptions of a specific incident where he was bitten by reduviid insects and from some of the symptoms he suffered later in life [115].

## 2.4.2 The Complex Life Cycle of *T. cruzi*

*T. cruzi* has a relatively complicated life cycle, with several distinct morphological stages, vector species and mammalian hosts (Figure 5). The life cycle begins in a *T. cruzi* reservoir, which is an infected animal or human [116]. Infected animals or humans have circulating parasites in the bloodstream. Reduviid insects consume a blood meal from the mammalian reservoir, taking up a population of *T. cruzi* trypomastigotes. Inside the insect, trypomastigotes pass into the midgut and differentiate to amastigote forms. Amastigotes are 3-5  $\mu\text{m}$  in diameter, proliferate and transform into epimastigotes in the midgut of the insect. There is no evidence that the parasite is harmful to the insect. Epimastigotes also proliferate, and move to the hindgut, where they transform to metacyclic trypomastigotes. Metacyclogenesis may be triggered by substrate interaction of the flagella [117]. Metacyclic trypomastigotes are

then excreted via insect feces, and infection can occur if feces come into contact with the bite wound or mucosal membranes. *T. cruzi* uses its flagella to move into the mammalian host.

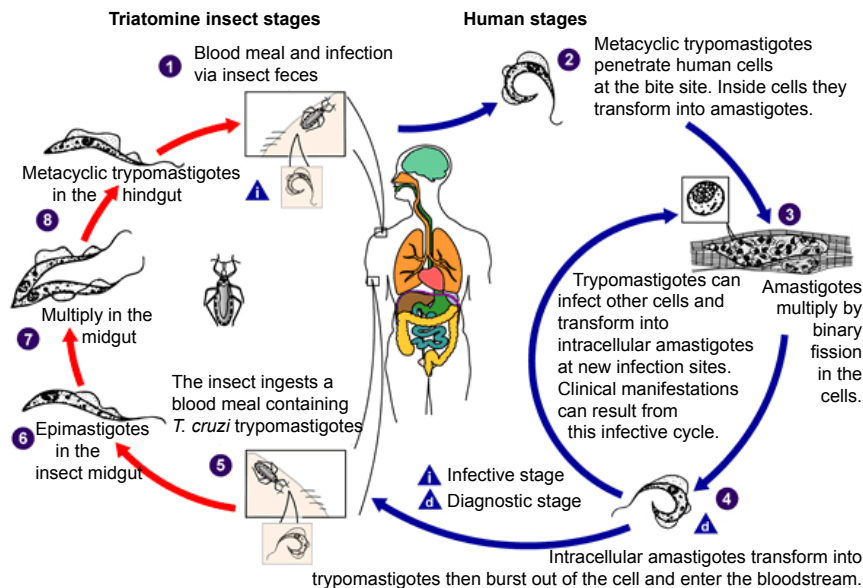


Figure 5: The *T. cruzi* life cycle. Image credit: U.S. Centers of Disease Control and Prevention.

Parasites then invade host cells via a mechanism involving the cytoskeleton and host cell lysosomes [118, 119]. Parasites are taken up by the lysosomes, and subsequent acidification inside the lysosomes activates parasite-secreted porin-like molecules that facilitate escape from the vacuole [120]. Once inside the cytoplasm, parasites differentiate into amastigotes and begin to proliferate, forming “pseudocysts,” and eventually turn into trypomastigotes again. Pseudocysts burst due to the parasite load, and large amounts of parasites are released to the bloodstream and they can infect new cells or get ingested by reduviid insects.

### 2.4.3 The Population Structure of *T. cruzi* is Wide, Complex and Contains Signatures of Ancestral Hybridization

The parasite causing Chagas disease, *Trypanosoma cruzi sensu stricto* (*s.s.*), is the type species of the subgenus *Schizotrypanum*. In addition to *T. cruzi s.s.*, the *Schizotrypanum* subgenus harbors approximately half a dozen other

trypanosome species, often referred to as *T. cruzi*-like species. Most *T. cruzi*-like species are restricted to bats (order *Chiroptera*), and are morphologically difficult to discriminate [121]. One of these bat-restricted organisms is *Trypanosoma cruzi marinkellei* (*T. c. marinkellei*), which was first characterized by Baker *et al.* in 1978 [122]. *T. c. marinkellei* is regarded as a subspecies of *T. cruzi* and is indigenous to South- and Central American bats. The human infective lineage, *T. cruzi s.s.*, should therefore be referred to as the nominate subspecies *Trypanosoma cruzi cruzi*. However, in this thesis the human infective parasite is simply referred to as *T. cruzi*, or *T. cruzi s.s.* when applicable. Lewis *et al.* estimated the divergence of *T. cruzi s.s.* and *T. c. marinkellei* at 6.51 million years ago using the *gpi* gene [123].

*T. cruzi* propagates predominantly via binary fission. However, Gaunt *et al.* created hybrid clones of distinct *T. cruzi s.s.* strains (*in vitro*), and thereby showed that *T. cruzi s.s.* has an extant capacity for genetic exchange [48]. The mechanism of genetic exchange is somewhat unusual, involving fusion of cells followed by genomic erosion to a diploid genome. Sexual events have likely shaped the current population structure of *T. cruzi s.s.*, but have been sufficiently rare to allow clonal propagation during long periods of time [124]. *T. cruzi s.s.* is currently partitioned into six discrete typing units (DTU; TcI-TcVI; Figure 8). Two of these, TcV and TcVI are the result of ancestral hybridization events from TcII and TcIII [125]. In addition to the six DTUs, there is one genotype identified only in Brazilian bats, TcBat [126]. The genetic heterogeneity of *T. cruzi s.s.* may explain differential clinical manifestations. The null hypothesis of neutral DTU subdivision with respect to Chagas disease severity can safely be rejected, but there is currently no definite correlation between DTU and disease outcome.

#### 2.4.4 Was the Ancestor of *T. cruzi sensu stricto* a Bat Trypanosome?

South American trypanosomes diverged from African trypanosomes after the break-up of Gondwanaland, and evolved parasitism independently [127, 128, 129]. While it is impossible to precisely date the divergence, estimates based on ribosomal RNA genes and biogeographical data, suggest that it occurred 90 to 100 million years from present [128, 130]. This implies that the divergence of the present day *T. cruzi* and *T. brucei* predated the origins of insect vectors and placental mammals. A phylogeny of the closest known relatives of *T. cruzi* is shown in Figure 6. *T. brucei* occurs exclusively in Africa and is transmitted by tsetse flies. In contrast to South American trypanosomes, *T. brucei* could

have co-evolved with primates and hominids for many million years. On the other hand, human presence in the Americas stretches no further than ~30,000 to 40,000 years from present [131]. Hence, *T. cruzi* can only have been in contact with humans for this period of time. An increase in human agricultural activities ~10,000 years ago was likely the first contact of the parasite with humans, and at that time most infections were likely to have been accidental. Parasite infections then gradually became more prevalent when human dwellings became infested with the insect as an extension of its natural habitat. Possibly, deforestation and an increase of agriculture facilitated the spread of insect vectors and thereby the disease.

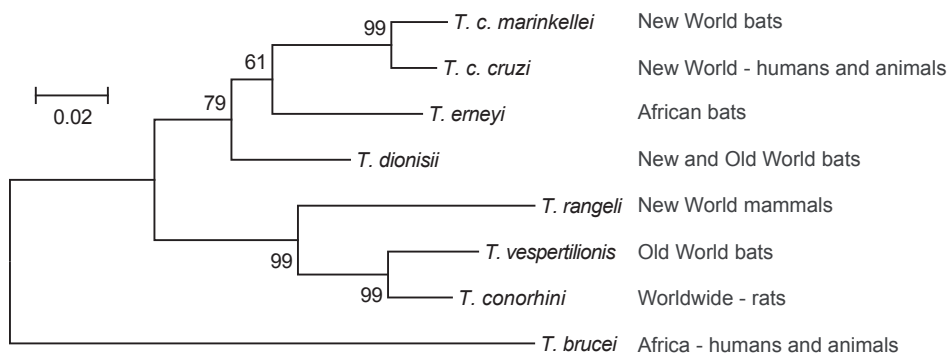


Figure 6: Maximum likelihood phylogeny of trypanosomatid species based on the *gGAPDH* gene. Alignments were done with ClustalW2 and inferred with MEGA5 [56] using the Tamura-Nei model, 1000 bootstrap replicates were performed. The scale bar refers to number of substitutions per site. Bootstrap values are close to the branches. *T. brucei* was used as an outgroup to support the phylogeny. Geography and hosts are indicated to the right. Accession numbers of the included sequences; JN040964, GQ140362, GQ140360, GQ140358, GQ140364, AJ620283, AJ620267, Tb927.6.4280 (GeneDB).

Current data indicate a strong association between bats and *T. cruzi*-like flagellates, suggesting a long period of shared evolutionary history. Thomas *et al.* reported that hematophagous arthropods might act as vectors for the transmission of *T. cruzi*-like species among bats [132]. Many of the infected bats are insectivorous, suggesting that bats become infected upon feeding on insect vectors. Hamilton *et al.* proposed the hypothesis that the ancestor of *T. cruzi* *s.s.* was a bat trypanosome, which made multiple jumps to terrestrial mammals [133]. Thus, the broad mammalian host range of *T. cruzi* *s.s.* may be a characteristic derived from a bat-restricted trypanosome. The



following observations support the “bat-seeding hypothesis” of *T. cruzi* s.s.: (i) the subgenus *Schizotrypanum* is dominated by bat-associated parasites; (ii) the closest known relative of *T. cruzi* s.s. is *T. c. marinkellei*, found only in South- and Central American bats; (iii) Lima *et al.* reported a new trypanosomatid species of African bats, *Trypanosoma erneyi*, forming a clade within the *Schizotrypanum* with *T. c. marinkellei* as a sister clade [134] (Figure 6); (iv) the present day *T. cruzi* s.s. has been found in bats, albeit at low prevalence [135, 136, 137]; (v) Marcili *et al.* recently reported a new genotype of *T. cruzi* s.s. (TcBat) that is only found in bats [126], which however little is known about and conclusions about its host specificity may reflect insufficient sampling; (vi) compared with *T. brucei*, the present day population structure of *T. cruzi* is wider and more complex, consistent with a dispersion facilitated by bats. One study recently reported new strains of the bat-restricted trypanosome *T. dionisii* in British bats, suggesting natural movement of bats between the Old and New World [138].

‘Ecological host switching’ describes a process where parasites may acquire new hosts or expand its host range without evolving new host utilization capabilities [139]. The process of ecological host switching has been proposed as the mechanism by which the ancestral *T. cruzi*-lineage colonized terrestrial mammals [140].

#### 2.4.5 An Unusual Amount of Genomic Redundancy

*T. cruzi* strains exhibit extensive variation in DNA content [141, 142, 143, 144], illustrating the diversity of this species. The genome sequence of the *T. cruzi* clone CL Brener (TcVI) has been determined [145]. The CL Brener clone is highly virulent and was isolated from the blood of mice infected with the parental strain CL [146]. The CL strain was originally isolated from *Triatoma infestans* in 1963 [147]. Several factors complicated genome finishing and resulted in a genome sequence of lower quality than those of *T. brucei* [148] and *Leishmania major* [149], which were sequenced in parallel: (i) the CL Brener clone was a genetic hybrid, i.e. it consisted of two 3-4% diverged haplotypes, referred to as non-Esmeraldo-like and Esmeraldo-like; (ii) the genome was enriched with sequence repeats of various types, comprising ~50% of the genome; and (iii) the karyotype of the CL Brener strain was found to be complex, consisting of at least 80 chromosomes of various sizes [150]. Arner *et al.* realigned the shotgun data from the genome project with the assembly and showed that many genes existed in almost identical copies [151]. This indicated that copy

number variation must have been introduced relatively recently, since alleles have had little time to diverge. Weatherly *et al.* organized contigs and scaffolds from the genome project into longer chromosome-wide sequences [150], although 30-40% of the genome is still unresolved. The majority of excluded genes belong to gene families. Most of the sequence repeats that went into the original genome draft [145] were genes encoding surface antigens, such as *trans*-sialidases (TSs), mucins, mucin-associated surface proteins (MASPs), dispersed gene family 1 (DGF-1), GP63 peptidases, and retrotransposons. Many of the repeated genes exist as pseudogenes in the genome. One prominent example is the TS family, containing at least 693 pseudogenes in the draft genome sequence, and likely many more alleles that fell outside the assembly [145]. Many of the surface proteins are glycosylated, and they cover and shield the parasite from the host immune system. Some TS proteins transfer sialic acid from the host to mucins [152], and have been proposed as drug targets [153]. Minning *et al.* used comparative genomic hybridization to sample multiple independent strains and found widespread copy number variation and whole chromosome aneuploidies [154].

Most *T. cruzi* genes are densely packed into polycistronic transcription units (PTU), which are separated by strand switch regions [103]. RNA polymerase (pol) II drives transcription in two different directions, resulting in polycistronic pre-mRNAs. The pre-mRNA from the PTU is subsequently matured via *trans*-splicing and polyadenylation to mRNAs. *trans*-splicing involves the ligation of a 39-nt spliced-leader sequence to the 5' end of transcripts. Since the life cycle is complex, the parasite needs to regulate its gene expression in order to adapt to different hosts and local environments. While RNA pol II is responsible for the overall transcription of PTUs, the genome lacks defined promoter elements. This organization suggests that individual genes are not regulated at the transcriptional level; rather it is assumed that gene expression is regulated at the post-transcriptional level. However, recent evidence indicates that epigenetic mechanisms may be involved: (i) Respuela *et al.* provided evidence of acetylation and methylation at divergent (head to head) strand switch regions, but did not find these patterns at convergent (tail to tail) strand switch regions or within PTUs [155]; and (ii) Ekanayake *et al.* reported the presence of the glycosylated thymine base ( $\beta$ -D-glucosyl-hydroxymethyluracil or base J) close to PTUs [156, 157]. Base J is rare outside of the *Kinetoplastida*, it has only been found in *Diplonema* (a small phagotropic marine flagellate) [158], and *Euglena gracilis* (a unicellular

algae and close relative to kinetoplastids) [159].

*T. cruzi* also contains mitochondrial DNA, which is present in a disk-like structure known as the kinetoplast. Kinetoplast DNA can be divided into maxicircles and minicircles and are circular molecules that are interlocked in a complex network. Their precise size depends on the strain, but maxicircles typically occur in 20-30 copies per cell and range in size from 35 to 50 kb. Minicircles occur in thousands of copies per cell and are approximately 0.8 to 1.6 kb. Transcripts of minicircles and maxicircles are involved in uridine insertion/deletion RNA editing [160]. Some evidence indicates that minicircles can integrate into the host genome [161, 162], and therefore have the potential to evoke immune responses and alter host gene expression.

#### 2.4.6 Lack of RNA interference in *T. cruzi* but not in *T. brucei*

Small RNAs are non-coding RNA molecules that are either functional or non-functional. In many eukaryotes, functional small RNAs are abundant and partitioned into many different classes, e.g. microRNAs, short interfering RNAs, piwi-interacting RNAs and several others [163]. RNA interference (RNAi) is a gene silencing process, deeply rooted in eukaryotes, mediating silencing via RNA-induced degradation of target transcripts. At the heart of RNAi lies the Argonaute/Piwi protein complex, which exerts post-transcriptional gene silencing. In the kinetoplastids, the presence of RNAi is variable [164]. Ngô *et al.* showed already in 1998 that *T. brucei* possesses functional RNAi [165, 166], but RNAi is missing or non-functional in *T. cruzi* [167]. In *Leishmania* spp. the situation is similar, RNAi is present in some species (e.g. *L. braziliensis*, *L. panamensis*, *L. guyanensis*) but not in others (e.g. *L. major*, *L. donovani*, *L. mexicana*) [164]. These data suggest that RNAi was lost twice in the evolution of the kinetoplastids. Future research will be needed to answer if *T. cruzi*-like organisms are also RNAi-negative. The cause of RNAi loss can only be speculative, but it is possible that loss of active mobile elements freed the parasite from keeping RNAi to mitigate the effects of mobile elements. It is also possible that loss of RNAi was selected for, i.e. if the loss altered gene expression so that it affected virulence or other properties.

While small RNAs of *T. brucei* have been relatively well studied [168, 169, 170], *T. cruzi* has received less attention on the subject. Analysis of the genome sequence has shown that *T. cruzi* lacks Dicer and Argonaute homologs [145]. However, the genome contains a gene encoding an Argonaute/Piwi-like protein, but apparently without a recognizable PAZ domain [171]. The

significance of this gene is uncertain. However, the lack of functional RNAi does not exclude the existence of small RNAs that may exert effects through other pathways. Small RNA species of *T. cruzi* containing the spliced leader mini-exon and other small RNAs were reported early [172, 173]. Garcia-Silva *et al.* reported a population of tRNA-derived small RNAs that localized to cytoplasmic granules, and increased during nutritional stress [174].



## 3 Aims

... we have come to the edge of a world of which we have no experience, and where all our preconceptions must be recast.

– D’Arcy Wentworth Thompson (1917), biologist.

The aim of the thesis was to further understand intraspecific genomic variation and transcriptional features of the protozoan parasites *Giardia intestinalis* and *Trypanosoma cruzi*.

### 3.1 Specific aims

#### **Paper 1**

Genome sequence comparison of the two human-infecting genotypes of *Giardia intestinalis* (A and B).

#### **Paper 2**

Identify genomic features that distinguish a non-human associated genotype of *Giardia intestinalis* (E) from two human-infective genotypes (A and B).

#### **Paper 3**

Genome comparison of *Trypanosoma cruzi sensu stricto* with the bat-restricted subspecies *T. cruzi marinkellei*.

#### **Paper 4**

Characterization of the short, non-coding transcriptome of *Trypanosoma cruzi* in order to understand if the parasite has functional classes of small RNAs.

#### **Paper 5**

Characterization of the transcriptome of *Giardia intestinalis* at single nucleotide resolution and investigation of gene expression divergence.



## 4 Present investigation

*Science is what we have learned about how not to fool ourselves about the way the world is.*

– R.P. Feynman (1918 - 1988), theoretical physicist.

### 4.1 Paper *i* and *ii*: Genome Comparison of Three Distinct Isolates of *Giardia intestinalis*

In Paper *i* and *ii* we performed genome comparisons of three distinct isolates (strains) of *G. intestinalis*, representing genotypes (syn. assemblages) A, B and E (see Figure 4 for a phylogeny of genotypes A to G). While A and B infect humans, the E genotype is only associated with hooved animals. The representative isolates are summarized in Table 1.

Table 1: Summary of compared isolates

Isolate	Assemblage	Host	Country	Ref. <sup>a</sup>	Accession <sup>b</sup>
WB	AI	Human	Afghanistan	[175]	AACB00000000.1 <sup>c</sup>
GS	B	Human	U.S.A.	[176]	ACGJ00000000.1
P15	E	Pig	Czech Rep.	[177]	ACVC00000000.1

<sup>a</sup> Original description of the isolate.

<sup>b</sup> NCBI GenBank accession number of the genomic data.

<sup>c</sup> An updated record (AACB00000000.2) is now available.

Morrison *et al.* described the genome of the WB isolate using Sanger sequencing [62]. In paper *i* and *ii* we sequenced the genomes of GS and P15 using Roche/454 sequencing (the former with FLX chemistry and the latter mainly with TIT chemistry; see [178] for an overview of the technology). Sequence assembly was performed *de novo* using the assembly software MIRA (Chevreux B, unpublished), and contiguous sequences (contigs) were then improved using targeted Sanger sequencing. Assembly and finishing resulted in 2,931 (N<sub>50</sub> 34,141 bp) and 820 (N<sub>50</sub> 71,261 bp) contigs of the GS and P15 genomes respectively. The assemblies were more fragmented than that of WB, but still represented the complete genomes of GS and P15, as confirmed by analysis of non-assembled reads and assembly sizes (11,001,532 bp of GS and 11,522,052 bp of P15). The P15 assembly was slightly more contiguous, due to the slightly longer read lengths of the TIT chemistry.

Contigs of the two genomes were annotated using a two-tier approach: (i) automatic transferring of gene models from the reference strain [62]; and (ii) followed by manual curation. The reference genome (WB) was downloaded from the database GiardiaDB [179], which is part of the EupathDB initiative to integrate genome-wide data sets from eukaryotic pathogens. Open reading frames (ORFs) were extracted from GS and P15, and annotated using best reciprocal BLAST toward genes of WB. Gene models were then manually inspected and unlikely gene models were discarded. Cross-genome comparisons allowed selection of the most conserved and therefore most likely start codon. The majority of the *nek* and *p21.1* families were assigned orthologs, but this was not the case for most *vsp* and *hcmp* genes. This suggested that *vsps* and *hcmps* have undergone lineage-specific diversification events, for example positive selection or recombination.

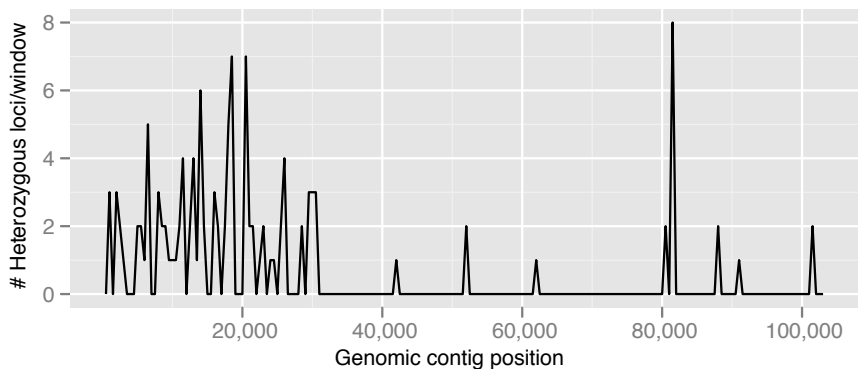


Figure 7: Heterozygous loci of the GS isolate counted in sliding windows along a genomic contig. Heterozygous loci were determined using alignments of Roche/454 reads. Windows were of the size 1000 bp and overlapped with 50%. The contig has the accession number ACGJ01002920. (Y-axis) Heterozygous loci per window. (X-axis) Position along the sequence.

Genomic heterozygosity was almost absent in the WB isolate ( $<0.01\%$ ) [62], the same was found in P15 with extremely few heterozygous loci. Conversely, GS exhibited extensive genomic heterozygosity, which was genome-wide estimated to  $\sim 0.5\%$ . The data did not *per se* reveal whether the observed differences were located in the same or different nuclei. Figure 7 shows how heterozygosity varies along a contig representing  $0.83\%$  of the genome. About half of detected heterozygous loci were located in coding sequences, and  $38\%$  changed amino acid (non-synonymous changes). As expected, there was a

strong bias toward transitions. One implication of a heterozygous genome is the potential to encode additional protein isoforms. Interestingly, heterozygous loci were frequently clustered, intervened by long homozygous regions. The presence of heterozygous loci in clusters rather than homogeneously dispersed over the whole genome are in disfavor of the Meselson effect, i.e. the phenomenon where asexual organisms accumulate heterozygosity in the absence of sexual or ameiotic recombination. A similar mosaic pattern was observed in *Naegleria gruberi* [180]. As seen in Figure 7, heterozygous loci are more predominant on the left half of this genomic segment. One interpretation of this pattern would be that homogenization via gene conversion has partially taken place. It is also possible that other processes have contributed to heterozygous deficit, such as the Wahlund effect or selfing/homogamy [181]. These indirect observations thus suggest that GS has undergone a more recent sexual event as compared with WB and P15. The two latter isolates may have longer asexual histories.

#### 4.1.1 The Core Genome and Isolate-specific Genes

The shared and non-shared gene content of the three isolates was investigated using reciprocal BLAST searches. The analysis revealed that the core gene content could be defined by 4557 genes, which excluded isolate-specific genes and *vsps*. The core gene content was comprised of housekeeping genes with homology to other eukaryotes, and *G. intestinalis*-specific genes. Thirty-eight, thirty-one and five genes were specific for P15, GS and WB respectively. One of the P15-specific genes represented an acetyltransferase, and phylogenetic analysis indicated a bacterial origin (likely from a bacterial species of the group Firmicutes). The donor lineage could not be precisely defined, but may be one of *Lactobacillus*, *Clostridium*, *Anaerotruncus* or *Enterococcus*, all of which are common inhabitants of the gastrointestinal tract. This suggests the uptake of the gene was relatively recent, and is an example of bacteria-to-eukaryote horizontal gene transfer. The GS genome contained several genes that were likely transferred from bacteria, one of which is likely an example of “dead upon arrival,” i.e. it was most likely transferred as a pseudogene. At least 96 genes with detectable homology to bacterial genes were conserved in the A, B and E genotypes and have attained housekeeping functions. The mechanism behind horizontal gene transfer is not determined, but likely involves multiple successive steps; where each must be successful in order for the gene to be integrated into the new genome. A successfully integrated



gene must also not be deleterious to the new host and convey a selective advantage to become fixed in the population. Frequent exposure to immense bacterial populations in the intestine has most likely contributed to creating opportunities for horizontal gene transfers.

#### 4.1.2 Structural Variation and Overall Divergence

Synteny breaks refer to disruption of gene order, often due to genomic rearrangements. In both the GS and P15 (compared to WB), synteny breaks were recorded and tended to occur in regions devoid of housekeeping genes. These regions displayed an atypical nucleotide composition in a sliding-window analysis and deviated in GC-content. Rearrangements may have been introduced spontaneously without affecting parasite fitness, and may therefore have circumvented purifying selection.

The average amino acid identity between WB and P15 was 90%, 81% between P15 and GS and 78% between GS and WB, as measured by comparing orthologous genes. The sequence identities recapitulated that of single-gene phylogenies, confirming the accuracy of previous phylogenetic trees, which are usually not inferred from genome-wide data. The sequence divergence of P15 and WB was similar to what is observed between *L. major* and *L. infantum*, whereas the divergence of GS and WB was similar to that of *Theileria parva* and *T. annulata*. Hence, the divergence between the studied *G. intestinalis* genotypes is similar to what is observed between distinct species. It can thus be argued that the *G. intestinalis* genotypes should be regarded as separate species rather than genotypes.

The dN/dS ratio (rate of non-synonymous nucleotide substitutions/rate of synonymous nucleotide substitutions) can be used to indirectly identify positive selection. The WB vs. GS comparison did not allow calculation of dN/dS ratios since synonymous changes were saturated (i.e. more than one substitution per site). Analysis of dN/dS of WB and P15 indicated, as expected, that most of the genome was under purifying selection. Several uncharacterized genes were found to exhibit elevated dN/dS ratios ( $>1$ ), indicating putative positive selection. Gene Ontology analysis indicated that five GO categories contained genes under positive selection, possibly reflecting co-evolution of multiple genes involved in common pathways. Furthermore, SAGE data were used to categorize genes into developmental categories. Four developmental categories displayed elevated dN/dS ratios, possibly reflecting lineage-specific divergence of cellular processes.

## 4.2 Paper *iii*: Genome Comparison of *Trypanosoma cruzi* *sensu stricto* With the Bat-restricted Subspecies *T. cruzi marinkellei*

The genome of the human infective *T. cruzi* (*T. c. cruzi*; *T. cruzi s.s.*) was compared with that of its closest relative, *T. c. marinkellei* (*Tcm*). Two clones were selected for the comparison: (i) *T. cruzi s.s.* Sylvio X10, which was isolated in 1983 from a human male in Pará State Brazil, and has confirmed pathogenicity [182]. Sylvio X10 is subgrouped into DTU TcI. (ii) *Tcm* clone B7, which was originally isolated in 1974 from the bat host *Phyllostomus discolor* in São Felipe Bahia Brazil (M.A. Miles and T.V. Barrett) [122]. *Tcm* B7 was not found to be infective in immunocompromised mice, nor did it provide immunological protection against subsequent challenge with *T. cruzi s.s.*, suggesting distinct antigenic profiles [122]. *Tcm* is restricted to South American bats and has to date not been recovered from humans. The phylogenetic relationship of *T. cruzi s.s.* and *Tcm* is shown in Figure 8.

*T. cruzi s.s.* Sylvio X10 (SX10) is a non-hybrid strain, and it has a smaller genome [144] than the previously sequenced *T. cruzi s.s.* strain CL Brener (TcVI; Figure 8) [145]. The smaller genome and non-hybrid nature implies that this clone likely has fewer repetitive sequences. *T. cruzi s.s.* SX10 is also evolutionarily distinct to *T. cruzi s.s.* CL Brener, and therefore creates an interesting basis for comparison.

Table 2: Summary of compared *T. cruzi* clones

Species	Subspecies	DTU <sup>a</sup>	Clone	Host	Ref.
<i>T. cruzi</i>	<i>T. c. cruzi</i>	TcI	Sylvio X10/1	Human	[182]
<i>T. cruzi</i>	<i>T. c. cruzi</i>	TcVI	CL Brener	Human	[147]
<i>T. cruzi</i>	<i>T. c. marinkellei</i>	-	B7	Bat	[122]

<sup>a</sup> Discrete Typing Unit; genotype.

The study began with Roche/454 and Illumina sequencing of sheared genomic DNA of *T. cruzi s.s.* SX10 and *Tcm* B7. Briefly, Roche/454 and Illumina reads were assembled separately *de novo* and the assemblies were then merged into a non-redundant assembly. The merged assemblies were subjected to quality enhancements, including scaffolding, gap closure and homopolymere error correction. The *T. cruzi s.s.* CL Brener genome was used for transferring gene models to the new genomes, since the majority of the gene repertoire was expected to be shared. The genomes were annotated us-

ing a semi-automatic pipeline. Orthologs were identified using best reciprocal BLAST, and gene models were manually curated. Flow cytometry estimated the genome sizes of *T. cruzi* s.s. SX10 and *Tcm* B7 to 44 and 39 Mb respectively (haploid genome sizes). The haploid genome size of *T. cruzi* s.s. CL Brener has previously been estimated to 55 Mb [145]. The assembly sizes of the genomes closely correlated with the experimental measures, which confirmed the computational steps.

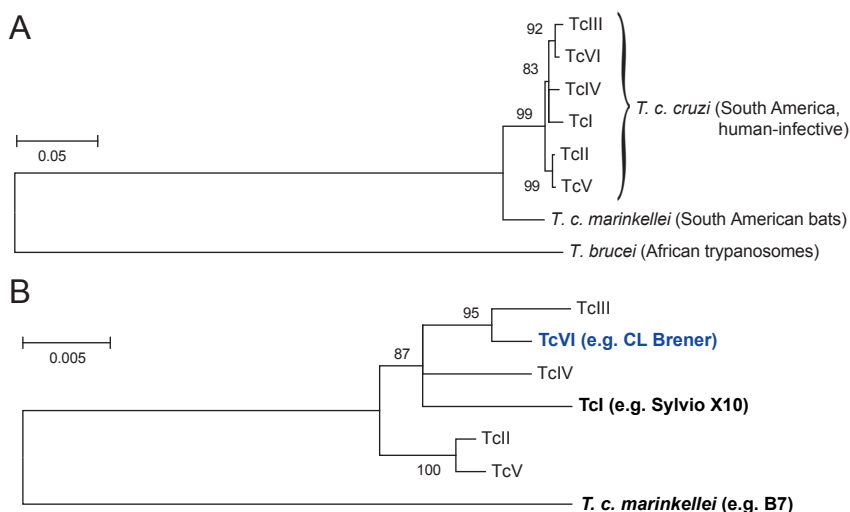


Figure 8: Nucleotide maximum likelihood phylogenies of *T. cruzi* s.s. DTU TcI-VI. The phylogeny was inferred from concatenated sequences encoding the beta-adaptin and endomembrane proteins. Sequences were aligned with ClustalW2 and inferred with MEGA5 [56] using the Tamura-Nei model. 1000 bootstrap replicates were performed. Scale bars refer to number of substitutions per site. Numbers close to branches indicate bootstrap support values. *T. brucei* (A) and *T. c. marinkellei* (B) were used as outgroups. Tip labels in bold indicate genotypes sequenced in the present study. (Blue tip label) Available reference genome (non-Esmeraldo-like haplotype). (Accession numbers) of the dataset: HQ859539, HQ859587, HQ859534, HQ859592, HQ859540, HQ859582, HQ859535, HQ859583, HQ859538, HQ859590, HQ859543, HQ859585, HQ859541, HQ859593, Tb927.10.8040, Tb11.02.0960.

Heterozygosity in the *T. cruzi* s.s. SX10 and *Tcm* B7 was estimated to 0.19 and 0.22% respectively. Sliding window analysis revealed that heterozygous sites were often clustered in blocks. The organization of heterozygous loci therefore resembled that found in *T. cruzi* s.s. CL Brener, although at much lower levels. One could speculate that the mosaic structure is a result of gene conversion. Overall, the heterozygosity levels were similar to those found in

some *Leishmania* species [183].

#### 4.2.1 Evidence of a Recent Eukaryote-to-Eukaryote Horizontal Gene Transfer

The genomes of *T. cruzi* s.s. SX10, *T. cruzi* s.s. CL Brener and *Tcm* B7 were systematically searched for gene differences. The search identified a unique 1,662 bp acetyltransferase gene in *Tcm* B7. The gene will be referred to by its locus tag MOQ\_006101. Sequence analysis revealed that the flanking genomic regions were present in *T. cruzi* s.s. SX10 and *T. cruzi* s.s. CL Brener, but not the gene itself. MOQ\_006101 was not identified in unassembled genomic reads of *T. cruzi* s.s. SX10 or *T. cruzi* s.s. CL Brener, suggesting it is unique to *Tcm* B7. Several fragments of VIPER retrotransposons were identified close to the locus, and RT-qPCR confirmed expression of the gene.

Phylogenetic reconstruction indicated that the closest known homologs were in algae and plants, suggesting MOQ\_006101 was transferred from another eukaryote. MOQ\_006101 showed low sequence identity (30-50%) toward genes in NCBI GenBank. The absence of intron-exon boundaries suggested MOQ\_006101 was transferred as a spliced mRNA, likely from a species not represented in GenBank.

The GC-content of the gene was compared with the global GC-content in coding sequences. GC-content analysis indicated an unusual composition compared to the rest of the genome, strengthening the notion of a transfer from another species. It remains to be determined whether MOQ\_006101 encodes a functional protein. Moreover, examination of multiple *Tcm* isolates may answer whether MOQ\_006101 has been fixed in the species or only in a certain lineage. Whether functional or not, the gene itself is interesting since it represents an unusual instance of horizontal gene transfer between two eukaryotes.

#### 4.2.2 *Trypanosoma cruzi marinkellei* has Capacity to Invade non-Bat Epithelial Cells

The potential of *Tcm* B7 to invade cell lines other than bat was investigated. Three common cell lines were selected: (i) Vero cells (kidney cells from African green monkey); (ii) OK cells (from a North American opossum); and (iii) Tb1-lu cells (bat lung). The experiments showed that *Tcm* B7 has retained the capacity to invade each of the three cell lines, despite that the cells were

originally derived from different species. Surprisingly, there was no preference for bat epithelial cells. Prolonged incubation of the parasite with cells showed that *Tcm* B7 is able to replicate intracellularly and the invasion process appears to be analogous to *T. cruzi s.s.*. In conclusion, these data indicate that the bat-specificity of *Tcm* is not mediated at the cell-invasion level.

### 4.3 Paper *iv*: The Small RNA Component of the *Trypanosoma cruzi* Transcriptome

A cDNA library of small RNAs (sRNAs) of the size range 16 to 61 nucleotides (nt) was prepared and sequenced from *T. cruzi* CL Brener epimastigotes, i.e. the insect stage of the parasite. Epimastigotes were used due to the ease of obtaining sufficient RNA from this life stage. Other stages often require cultivation together with mammalian cells, which may contaminate the subsequent RNA preparation. We also assumed that functional sRNAs, if present at all, would also be present in epimastigotes. The particular size range of 16 to 61 nucleotides was selected to avoid spliced leader RNA, which could otherwise cloud the analysis.

Sequencing generated 582,243 sRNAs, of which 90.7% aligned with the genome sequence. We subsequently used the annotation of the genome to assign sRNAs into relevant categories, i.e. if an sRNA overlapped a tRNA, it was assigned to the tRNA category, etc. With respect to the non-Esmeraldo-like haplotype, 72.1% of the sRNAs derived from transfer RNAs (tRNA-derived small RNAs; tsRNAs). 97.4% of sRNAs were derived from only three classes of non-coding RNAs (transfer RNA, ribosomal RNA and small nuclear RNA). Only 0.18% of sRNAs were derived from protein-coding genes. 2.42% of sRNAs could not be grouped into any canonical RNA class. A few of the small RNAs were experimentally validated using a real time quantitative PCR-based assay. In conclusion, the bulk of sRNAs of the 16 to 61 nt size range were derived from known classes of non-coding RNAs.

The median length of tsRNAs was 38 nt, and 88.9% of them were derived from the tRNA 3' end. One example of a tRNA-derived small RNA is shown in Figure 9. 75.3% of the 3'-derived tsRNAs contained the post-transcriptional 'CCA' extension, a hallmark of mature tRNAs. This indicated that most tsRNAs were derived from mature tRNAs, albeit not all. However, it is possible that the 'CCA' extension has been lost during sample or library preparation. If tsRNAs would represent degradation products of tRNA turnover, one would expect a correlation between sRNA copy number and the expression level of tRNA genes. Analyses of amino acid usage as a substitute of tRNA gene expression data did not find any correlation. The cleavage site of tRNAs was present inside the anticodon loop, suggesting endonucleolytic cleavage as the responsible mechanism of generation.

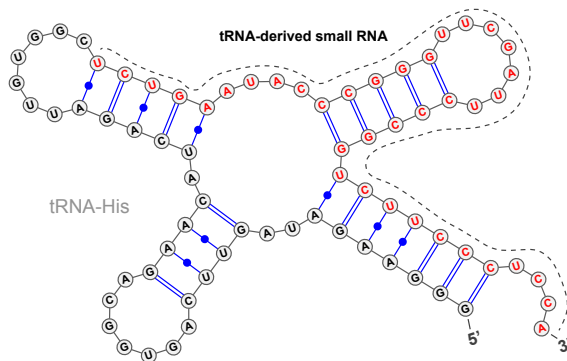


Figure 9: Secondary structure of tRNA-His from *T. cruzi* CL Brener (Tc00.1047053508861.10). The secondary structure was predicted using tRNAScanSE [187] and visualized using VARNA [188]. The sequence in red indicates the part cleaved into a small RNA with copy number 41,929 in the data set.

1.69% of the small RNAs were not derived from known non-coding RNAs. These small RNAs were clustered based on their genomic alignment coordinates. Clustering formed 92 distinct expression loci, of which homology searches revealed known non-coding RNAs for 13 loci. The remaining 79 loci did not fall into known non-coding RNA classes and had an average length of 54 nt. No homology was found in Rfam or GenBank databases. 35 of the novel RNAs folded into non-hairpin secondary structures and 18 folded into hairpin structures. 1,159 small RNAs were not clustered and had a median length of 24 nt. Of these small RNAs, BLAST searches revealed 335 to be derived from rRNA and 819 from protein-coding genes. MicroRNA target site prediction of the latter population using the “seed region” (nt 2-8) predicted the possibility that these regulate certain categories of mammalian genes. It is therefore tempting to speculate that the parasite may use small RNAs for inter-cell communication, or possibly modification of the gene expression response of the host cell.

0.13% of small RNAs were derived from repeats, including retroelements. In particular the CZAR element contained 446 mapped small RNAs, which may indicate putative initiation fragments from the transcription of these elements. There was no overrepresentation of antisense reads in any repeat class, suggesting that small RNAs have no role in perturbation of mobile elements. Searches did not reveal any canonical classes of regulatory sRNAs found in metazoans. This is consistent with the lack of RNA interference.

#### 4.4 Paper *v*: Transcriptome Profiling at Single Nucleotide Resolution of Diverged *G. intestinalis* Genotypes Using Paired-end RNA-Seq

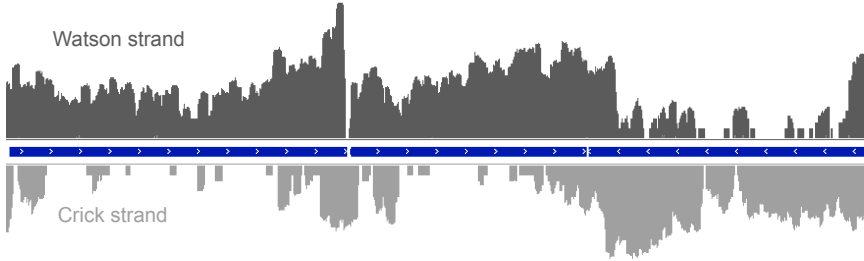


Figure 10: Coverage of RNA-seq sequence reads on three open reading frames (logarithmic scale; blue horizontal bars: GL50803.92132, GL50803.92134, GL50803.40369). (Watson strand; “plus”) Top. (Crick; “minus”) Bottom. The strand of the open reading frame is shown by arrows (‘>’ refers to the plus strand and ‘<’ refers to the minus strand).

In this study we performed transcriptome sequencing (RNA-Seq) and comparison of the polyadenylated transcriptomes of four diverged isolates of *G. intestinalis*. The specific aims of the study were to: (i) identify genotype-specific patterns of gene expression; (ii) confirm and refine genome annotations; and (iii) identify qualitative transcript properties like 3′ untranslated regions (UTRs). Total polyadenylated RNA was harvested from *in vitro* grown trophozoites of the four isolates WB (AI), AS175 (AII), P15 (E) and GS (B). Sequencing libraries were prepared according to a strand-specific paired-end protocol and sequenced on Illumina HiSeq 2000 as 2x100-nt reads. Each library was sequenced as two technical replicates in order to estimate technical variation. The reproducibility of the RNA-Seq method was determined using two biological replicates of AS175. RNA-Seq generated 33 to 41 million read-pairs from each library, which were aligned (mapped) to the corresponding reference genome (Table 3). Figure 10 shows the strand-specificity of the obtained data. The aligned RNA-Seq data were then used to calculate digital gene expression values, formulated as fragments per kilobase of transcript per million fragments mapped (FPKM). 49 genes were selected for RT-qPCR validation, which confirmed the accuracy of the measurements. Moreover, a global comparison was performed with three microarray data sets from GiardiaDB [179], indicating a moderate but significant correlation of the two tech-



niques. RNA-Seq measurements did not correlate with SAGE data, which is not surprising since SAGE is generally not quantitative.

Table 3: Summary of studied isolates and generated RNA-Seq data

Isolate	Assemblage	Ref.	#Mapped <sup>b</sup>	%ORFs <sup>c</sup>	%Orthologs <sup>d</sup>
WB	AI	[62]	37,888,422	93.7	99.5
AS175	AII	<sup>a</sup>	36,878,839	98.3	99.8
P15	E	[189]	36,437,138	96.3	99.6
GS	B	[190]	31,869,806	97.3	99.7

<sup>a</sup> The genome sequence is not published.

<sup>b</sup> Reads of this strain that uniquely mapped to the reference genome.

<sup>c</sup> Percentage of [annotated] ORFs with detectable transcription (FPKM>0.5).

<sup>d</sup> Percentage of conserved four-way orthologs with detectable transcription.

#### 4.4.1 Low Biological and Technical Variation

Technical variation consists of measurement imprecision introduced during library preparation or by the sequencing instrument. The technical replicates of each sequencing library indicated very low technical variation (Pearson's  $r^2=0.99$ ). In this study technical replicates were subjected to the same library construction procedure, and thus reflected variation of the sequencing instrument. On the contrary, biological replicates reflect both technical and biological variation. Biological variation may result from uncontrolled environmental cues or from stochastic variation in gene expression. Biological variation was low ( $r^2=0.97$ ), and we therefore concluded that gene expression measurements were reproducible. The number of genes involved in the culture-induced biological response was estimated with a  $\chi^2$  test, and found to be 4% of the total gene content at  $p=0.01$  (AS175). There was no meaningful way to group the implicated genes.

#### 4.4.2 Gene Expression Levels Recapitulate the Known Phylogeny

Almost the entire *G. intestinalis* genome was transcribed to some extent, and provided transcriptional evidence for >99% of the conserved genes (Table 3). Transcription levels exhibited a wide dynamic range; the fold difference of the median of the 5% lowest and highest expressed genes was 873X. Notably, a transcriptionally silent gene cluster was identified on chromosome 5 of the WB isolate, encompassing 28 genes in tandem (Figure 11). The genomic region was 41 kb and contained genes associated with replication and genes of no

known function. Sequence signatures of these genes were also identified in the GS isolate, suggesting that the region was acquired prior to the split of the lineage leading to the extant A and B genotypes. Because the region exhibited higher than average divergence between A and B, it is possible that it has been subjected to genetic drift without purifying selection. Saraiya *et al.* reported one ORF of this cluster to transcribe a microRNA-like small RNA [81], suggesting that the region may have certain functionality and provides an explanation to why it has not been lost. Nevertheless, the lack of detectable transcription suggests it may be a selected feature, possibly due to negative effects of parasitic DNA on fitness.

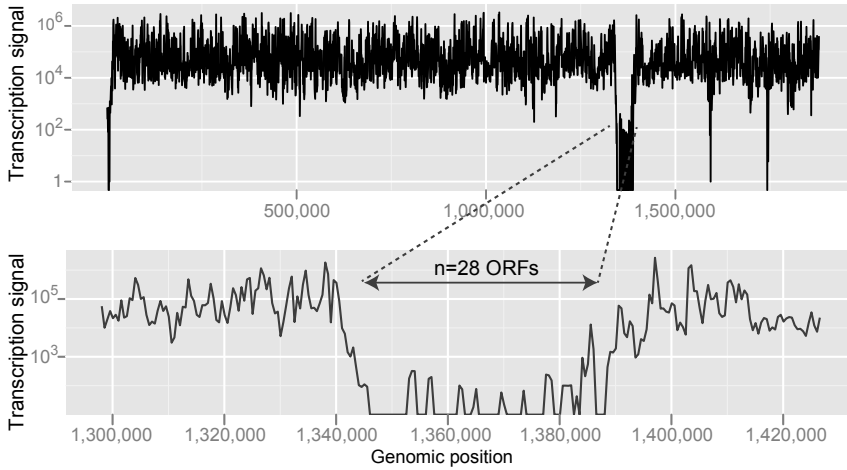


Figure 11: Sliding window analysis of RNA-Seq coverage on scaffold CH991767 (WB isolate). The zoomed in region shows drop in RNA-Seq coverage along a 41 kb region. Sequence coverage was calculated in 500 bp non-overlapping windows. (Y-axis)  $\log_{10}$ -transformed coverage sum. (X-axis) Position along scaffold.

Global gene expression profiles of the isolates were compared and genotype-specific gene expression was identified using a  $\chi^2$  test. A relatively limited number of genes were differentially expressed (31 to 145 genes). These numbers likely represent the lower detection bound since the implemented analysis was conservative. It remains to be investigated how many of the differentially expressed genes are of functional importance.

The theory of neutral evolution states that most nucleotide changes are randomly introduced by neutral drift without affecting fitness [192]. On the contrary, positive selection is driven by advantageous mutations and is gen-

erally more rare. It is currently accepted that coding sequences (CDS) predominantly evolve by neutral evolution followed by purifying selection [193]. Less is known about the modes of evolution acting on gene expression. We investigated if the rate of gene expression divergence was correlated with that of the CDS. CDS divergence was estimated by the rate of non-synonymous nucleotide substitutions (dN), while gene expression divergence was estimated by FPKM fold change. The analysis was performed genome-wide on the maximum number of defined ortholog pairs (the precise number varied slightly between isolates). Cross-correlations of any two isolates resulted in Pearson's  $r$  ranging from 0.069 to 0.11, indicating a weak correlation of the two variables. Conversely, there was no correlation between the rate of synonymous nucleotide substitutions (dS) and gene expression divergence (Pearson's  $r=0$ ). Interestingly, highly expressed genes tended to have lower rate of divergence (dN). These data indicated a limited coupling between CDS and gene expression divergence, suggesting that random drift has been the predominant evolutionary mode of gene expression divergence in *G. intestinalis*.

#### 4.4.3 Promiscuous Polyadenylation Sites and Unusual *cis*-acting Signals

Polyadenylation sites (polyA sites) were precisely mapped using RNA-Seq data, which allowed global analysis of the *G. intestinalis* 'polyadenylation landscape'. PolyA sites were mapped using reads (polyA tags) containing the mRNA:polyA junction (Table 4). Aligned polyA tags revealed 22,221 to 49,027 distinct polyA sites (depending on the isolate; Table 4).

Table 4: Mapped polyadenylation sites

Isolate	#tags <sup>a</sup>	#sites <sup>b</sup>	#PACs <sup>c</sup>	#transcripts <sup>d</sup>	Median 3' UTR (nt)
WB	456,928	49,027	7,617	3,884	80
AS175	183,454	37,028	5,028	3,057	100
P15	436,720	51,499	8,037	3,800	83
GS	71,118	22,221	2,624	1,651	85

<sup>a</sup> PolyA-tags with mapping quality >40.

<sup>b</sup> Unique polyadenylation sites.

<sup>c</sup> PolyA sites within 10 nt were clustered into polyadenylation site clusters. The numbers refer to number of clusters with  $\geq 4$  tags.

<sup>d</sup> Transcripts with an associated polyA site.

Each of the polyA sites represents the 3' end of an independent transcript, although not necessarily an mRNA. PolyA sites were found to exhibit mi-

croheterogeneity, i.e. imprecision of a few nucleotides. Microheterogeneity of polyA sites has been documented in higher eukaryotes and is attributed to the imprecise nature of the polyadenylation machinery [194, 195]. The phenomenon is not to be confused with alternative polyadenylation, which is regulated by distinct polyadenylation signals. To account for this heterogeneity, polyA sites within 10 nt were clustered into polyadenylation site clusters (PAC). PACs were assigned to the most likely ORF based on proximity, i.e. the 5'-most PAC counted from the translational stop codon was assigned to the ORF. Cloning and 3' rapid amplification of cDNA ends was performed on 9 genes for validation, confirming the accuracy of the mapped sites. The median 3' UTR length was found to be 80 nt for WB, and similar for the other isolates. In comparison, the median 3' UTR length of *Saccharomyces cerevisiae* is 104 nt [196]. This is longer than earlier estimates (around 30 nt) from a small set of highly expressed genes. Several microRNAs have been identified in *G. intestinalis* but searches for target sequences were only done within the first 50 bp from the stop codons. These results suggest that regulation of gene expression via microRNA binding to 3' UTRs can be common since many mRNAs have relatively long 3' UTRs.

The mapped polyA sites were analyzed for putative *cis*-acting signals, i.e. polyadenylation signals (PAS). Positions -40 to -1 of each polyA site were searched for overrepresented hexamers using an iterative algorithm described by Beaudoin *et al.* [197]. The search identified 13 prominent hexamers, which represent putative PAS. None of these were identical with the canonical eukaryotic PAS (AAUAAA). However, 5 of the 13 putative *G. intestinalis* PAS contained the tetranucleotide 'UAAA', which is a part of the canonical eukaryotic motif. The tetramer was located approximately 10 nt from the polyA site. Interestingly, polyA sites located in the sense orientation tended to have fewer of the identified hexamers in contrast to antisense or intergenic polyA sites.

The nucleotide composition surrounding polyA sites was analyzed, which indicated a distinct pattern of AU-richness. This pattern may be required for recognition by the polyadenylation machinery, or for binding of polyadenylation factors. 81% of tail-to-tail gene pairs had 3' UTRs that overlapped the transcription unit of an adjacent gene on the opposite strand. This transcriptional organization may be a way of gene regulation, but also causes the production of double stranded RNA. Such pervasive transcription is often not compatible with functional RNA interference. While *G. intestinalis* has some

components of RNAi, it is not completely identical to that found in higher eukaryotes. It can therefore be speculated if *G. intestinalis* has lost certain components of RNAi in favour of genomic fidelity and minimalism.

#### 4.4.4 Biallelic Transcription and Correlation of Allele Expression and Allele Dosage

As described in paper *i*, the GS isolate contains ~0.5% genome-wide heterozygosity. The RNA-Seq data were used to confirm expression of the identified alleles and to study allele-specific expression (ASE). Mapping bias is a major problem in ASE assays, which means that current mapping algorithms preferentially map one of the alleles (discussed in for example [14]). To further understand the extent of mapping bias, we generated and mapped simulated RNA-Seq data. The simulated data followed realistic error profiles, and did not indicate systematic mapping bias, but nevertheless indicated an inherent bias toward mapping of one allele. The amount of bias was influenced by sequence errors, but likely also other factors.

The simulated data was modeled using a Cauchy distribution and used for significance testing. 98% of the genes with at least one heterozygous locus displayed biallelic transcription, i.e. the two alleles were identified at the transcriptional level. Of these genes, 82% indicated allelic expression imbalance at  $p=0.05$ , i.e. not an equal number of reads were derived from each allele. We examined if the allelic expression ratio corresponded to the observed allele count. Allele counts were inferred from the depth of genomic reads. For the vast majority of heterozygous loci there was a linear correlation between the RNA-Seq signal and the inferred allele count. When only heterozygous sites corresponding to the allele ratio A:A:B:B were investigated, 59% of the analyzed sites displayed expression imbalance. It can be assumed that most of the allelic imbalance is caused by allele dosage rather than *cis* or *trans* regulatory differences. In conclusion, the current data indicate that both nuclei are transcriptionally active, and there was a correlation between expression level and allele dosage. Together these data indicate that transcription in *G. intestinalis* is symmetric.

#### 4.4.5 Only Six Genes are *cis*-spliced

Only five genes are currently reported to contain introns and undergo *cis*-splicing in *G. intestinalis*. These genes are listed in Table 5. We performed

an exhaustive search for new *cis*-splicing events using a comparative transcriptomics approach. Putative splice junctions of WB, AS175, P15 and GS were first mapped with the software TopHat [198]. Manual inspection of the deduced splice junctions indicated a large number of dubious suggestions, as concluded from the proposed splice pattern and the annotation of the involved genes. For example, many multicopy genes were suggested to undergo extensive splicing, e.g. *vsp* and *p21.1* genes. The repetitive nature of these genes is likely to give artifacts when mapping short reads, especially since the implemented algorithms chop reads into even smaller pieces before mapping them. To increase the signal to noise ratio, the algorithmically identified splice sites were filtered according to these criteria: (*i*) the splicing pattern was required to be conserved in at least two isolates; (*ii*) repeated genes were discarded (e.g. *vsp*, *hcmp* and *p21.1*); (*iii*) the intron had to be confined to the ORF or the closest upstream intergenic region; and (*iv*) the splice pattern had to be supported by a minimum of 5 reads.

Table 5: Confirmed introns in *G. intestinalis*

Gene	ID <sup>a</sup>	Ref.	Boundary <sup>b</sup>	nt <sup>c</sup>	#Isolates <sup>d</sup>	Pos. <sup>e</sup>
[2Fe-2S] ferredoxin	27266	[63]	CT-AG	35	4	0.05
<i>Rpl7a</i>	17244	[64, 62]	GT-AG	109	4	0.54
dynein light-chain	15124	[62]	GT-AG	32	3	0.05
uncharacterized	35332	[64]	GT-AG	220	2	0.01
uncharacterized	15604	[65]	GT-AG	29	4	0.02
uncharacterized	86945	novel	GT-AG	36	4	0.02

<sup>a</sup> Prefix: GL50803\_

<sup>b</sup> The intron boundaries, 5' to 3' (splice sites).

<sup>c</sup> Length of the intron in nucleotides.

<sup>d</sup> Number of isolates the splicing pattern was found in.

<sup>e</sup> Intron position in the gene along the 5' to 3' axis. The position was calculated as: [position of first nucleotide of the intron]/[gene length].

The five previously reported introns were found in our data, indicating that the bioinformatics procedure was robust. Three of the previously confirmed splice variants were found in all 4 isolates, and two were found only in 3 and 2 isolates, respectively (Table 5). This reflects limitations in our data sets or genome assemblies rather than differential splicing between isolates. The bioinformatic search suggested 14 new intron candidates. However, PCR validation on genomic DNA and RNA (cDNA) subsequently rejected 13 of the 14 candidates, which suggests that splice site prediction using RNA-Seq is noisy. One new intron was confirmed by PCR, and the amplified DNA was

sequenced with dye-terminator sequencing, confirming splice sites.

The novel intron was 36 nt in length and was present in an uncharacterized gene on chromosome 4 (Table 5; Figure 12). The intron was identified in all four isolates and had canonical splice boundaries (GT-AG). Removal of the intron extends the open reading frame with 73 codons. It is possible that some putative introns were missed, especially low-level *cis*-splicing events. However, the true number of introns in *G. intestinalis* is not likely to be much higher than presented here.

```

atgctggattctgtgatctctctttttcttgcgcccttcgcgaagaaggtgtaccagaa
M L D S V I S L F L A A L R E E G V P E
      10      20      30      40      50      60
gccccaaacactcgagctgctccagaccatccgttcctggccacagatcaaggcaagcaca
A Q T L E L L Q T I R S W P Q I K A S T
      70      80      90     100     110     120
tatactatcgtatgatttattttttcccaacagcctaacacacagatacagacacttaat
Y T I                                I Q T L N
      130     140     150     160     170     180
aaccttgctaccagaggagtagcagagcgtccaaagcgcttacagatattaccaccacattt
N L A T R G V R A S K A L T D I T T T F
      190     200     210     220     230     240
ttcacttctccgcgaatgtctacagcgcaggggtctttcttgtcaagacctagatgctttt
F T S P R M L Q R R G L S C Q D L D A F
      250     260     270     280     290     300
catgacttttagtggcgtgattgtaagaaattttattgtccatgggcatcagatccatggg
H D F S G V I V R N F I V H G H Q I H G
      310     320     330     340     350     360
gttggttttactcctcttcagcttcttaga
V G F T P L Q L L R
      370     380     390

```

Figure 12: 36-nt intron (underlined) in the gene GL50803.86945, encoding an uncharacterized protein. Only the first 390 nt of the gene is shown. The translational start codon is indicated in green and the conceptual translation is shown in blue.

Five out of six introns displayed a positional bias towards the 5' end of the gene (Table 5). Intron deficit and 5' bias have been observed in the two protozoa *Encephalitozoon cuniculi* [199] and *Guillardia theta* [200], both of which belong to intron-rich groups of organisms. Gene sampling of species of the order *Oxymonadida* (anaerobic flagellates found mainly in the gut of termites and wood-eating roaches) have found evidence of protozoa with extensive intron content [201]. These data corroborate the idea that *G. intestinalis* derived from a more intron-rich ancestor but lost introns during evolution.



## 5 Concluding remarks

*There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.*

– C. Darwin (1809-1882), *On the Origin of Species*.

### 5.0.6 Comparative Genomics

The sequence data generated in the present studies are publicly available in integrative databases together with data sets from independent studies. Much of the data have already been utilized in several independent studies (for example [67, 202, 203, 204, 205]), which underscore the value of large genomic data sets in parasitology: both to increase understanding of parasite biology and for hypothesis generation. As sequencing technologies continue to improve in terms of output and costs, it will be of value for the research community to undertake large-scale efforts to sequence multiple genomes of biologically relevant strains. Similar to the 1000 genomes project [206], future efforts in parasitology may target several hundred or more strains. The integration of such data sets into community-oriented databases, for example EuPathDB [8], should allow researchers to take advantage of the vast amount of data. One challenge for the future will be to develop suitable phenotyping strategies for protozoan parasites, since sequence data from many strains may be of limited value if phenotypes are unknown. Second, we can learn about the evolutionary adaptations that led to parasitism from genome comparisons of even more diverged species, for example free-living or avirulent species within the diplomonadida and kinetoplastida. Examples of such species are *Spironucleus* sp. (diplomonadida) and *Trypanosoma rangeli* (kinetoplastida). Such sequencing projects are currently being undertaken and are likely to yield new insights into parasite evolution.

Many genes of these parasites are currently uncharacterized. A future priority of parasitology should therefore be to explore the functionality of uncharacterized genes, since these are the genes likely to be parasite-specific rather than universally conserved among eukaryotes. In theory, it would also



be possible to find drug targets among these genes. In conclusion, many questions relating to the basic biology and evolutionary history of these parasites are still unresolved. For example, do isolate-specific genes contribute to the phenotype? What is the mechanism behind horizontal gene transfer and how often does it happen? Why do some parasite lineages contain higher heterozygosity than others? Do protozoan parasites recombine and how often? Genome sequencing may not by itself provide answers to these questions; rather genomic data should be used together with functional studies or other data sets. Ultimately, this may facilitate new insights into the basic biology that may lead to new treatments and control of these pathogens.

#### 5.0.7 The Short Transcriptome of *T. cruzi*

There was no evidence of canonical small RNAs as often found in metazoans, e.g. microRNAs, small interfering RNAs or piwi-interacting RNA. This finding is consistent with the lack of RNA interference in *T. cruzi*. More than 90% of small RNAs were derived from known non-coding RNAs (tRNA, rRNA and snRNA). The most prominent category was tRNA-derived small RNAs. It remains to be determined if these small RNA classes are functional, partially functional or merely debris from the RNA “degradome,” i.e. turnover. The following observations warrant further investigation of *T. cruzi* small RNAs: (i) the cleavage pattern appears to be non-random; (ii) certain small RNAs appear to be stable, as suggested by the copy number; (iii) tsRNA locates to cytoplasmic granules [174]; and (iv) certain tRNA isoacceptors were over-represented in terms of deriving tsRNAs. Interestingly, we identified a population of small RNAs that were not derived from known non-coding RNAs, and were predicted to contain microRNA-like seed regions. The present study only briefly scratches the surface of the small RNA transcriptome and raises questions that call for further investigation.

#### 5.0.8 Comparative Transcriptomics of *G. intestinalis*

Transcription is one key event in the translation of genotype to phenotype, and transcriptome studies can enhance our understanding of phenotypic differences and the evolutionary trajectories of pathogens. Almost the entire genome of *G. intestinalis* was transcribed in trophozoites grown *in vitro*, confirming the promiscuous nature of transcription in this parasite. The data confirmed many gene models that were originally annotated without tran-

scriptional evidence, and also identified novel protein-coding genes. The extent of gene expression divergence recapitulated the known phylogeny of the strains, suggesting that gene expression has largely evolved via genetic drift. Despite many genes with strain-specific expression, it is difficult or impossible to conclude what differences are of functional importance. Furthermore, it is important to note that differences in mRNA abundance may not necessarily result in differences at the protein level.

Notably, a non-expressed gene locus was identified, consisting of 28 non-transcribed genes, which may reflect transcriptional repression by a yet-to-be described mechanism. Sequence signatures indicated a putative viral origin. Perhaps silencing operates at the level of chromatin organization. Functional investigation will be required to elucidate the mechanism behind this mode of silencing.

Biallelic transcription was identified at most of the heterozygous loci of the *G. intestinalis* isolate GS, which were identified and described in paper *i*. Comparison of allele dosage with allelic transcription levels indicated a relationship between the two variables. The data corroborate previous observations of binucleic transcription in *G. intestinalis* and further provide an association between allele dosage and transcription levels. These data suggest that binucleic transcription is largely symmetric, which nonetheless does not preclude the existence of allele-specific expression. However, the latter does not appear to be a general feature of transcription in *G. intestinalis*, which is also consistent with the deficit of regulation at the transcriptional level.

Previously reported introns were confirmed and only one new intron was discovered, suggesting that the true number of introns is not likely to be much higher than this. It is tempting to speculate that *G. intestinalis* may have been more intron-rich in the past, but undergone intron-loss. Conversely, it is also possible that introns became more prevalent in eukaryotes after the split of *G. intestinalis* from the main eukaryotic lineage. However, it seems unlikely that six introns have necessitated the evolution of the relatively complex splicing machinery of *G. intestinalis*. The final evidence to settle this question would be the finding of a diplomonad relative with an extensive repertoire of introns, which is yet to be discovered.



---

## 6 Populärvetenskaplig sammanfattning

De två parasiterna *Trypanosoma cruzi* och *Giardia intestinalis* är encelliga organismer som orsakar sjukdom och lidande hos flera miljoner människor världen över. *Trypanosoma cruzi* är huvudsakligen ett problem i Latinamerika där den ger upphov till Chagas sjukdom. ~8 miljoner människor är infekterade i Latinamerika och 11,000 avlider varje år till följd av sjukdomen [97]. I Sverige finns grovt uppskattat 1,000 infekterade personer [207]. *Giardia* drabbar människor runt om i hela världen och kan leda till diarrésjukdomen giardiasis. Migration leder till att sjukdomarna blir vanligare i Europa, Nordamerika och andra delar av världen. Både Chagas sjukdom och *Giardia* räknas som försummade sjukdomar, och orsakar problem i utvecklingsländer [26, 97]. De nämnda sjukdomarna, tillsammans med flera andra tropiska sjukdomar, prioriteras ofta inte i fråga om läkemedelsutveckling eftersom det inte anses ekonomiskt lönsamt.

I den här avhandlingen har jag använt datoranalyser för att studera biologisk information från de två nämnda parasiterna, bland annat gener som kodas i dess arvsmassa. Information från parasiterna avläses med särskilda instrument som med hög noggrannhet avläser den genetiska koden. Därefter analyseras informationen för att hitta biologisk relevanta mönster som kan öka förståelsen av parasiternas biologi. Detaljerade analyser och jämförelse mellan olika stammar beskrivs, och identifierar egenskaper som är kodade i parasiternas DNA. De metoder som använts avslöjar också evolutionära mönster, som kan bidra till att öka förståelsen för hur parasiterna anpassat sig till människan och hur de undviker immunförsvaret. I den här avhandlingen har även genernas aktivitet studerats, det vill säga genernas uttryck. Genuttrycket kan avslöja hur olika stammar skiljer sig funktionellt. Information från de genomförda studierna kan i framtiden användas för att designa mer skraddarsydda experiment som kan leda till bättre behandlingsmetoder och diagnostik.

---

## 7 Acknowledgments

*Skepticism is the agent of reason against organized irrationalism – and is therefore one of the keys to human social and civic decency.*

– S.J. Gould (1941-2002), evolutionary biologist.

I'm grateful to have been given the opportunity to do a PhD and for the freedom to be creative. I want to thank and acknowledge friends and colleagues I met during my PhD studies, without whom the journey would have been less fun and fulfilling.

My supervisor prof. **Björn Andersson** – for scientific and moral support and for also allowing me to learn science abroad. My co-supervisor prof. **Staffan Svärd** and his group – for good science, support and travel company.

In addition, I acknowledge the following people (the names are not in any particular order): **Stephen Ochaya** – for chats about science and life. **Jon Jerlström-Hultqvist** – for experiments, discussions and collaboration, and for travel company to Kampala and Wellington among others. **Marcela Ferella** – for experiments, discussions and collaboration. **Lena Åslund** – for discussions on *T. cruzi* small RNAs. **Erik Arner** – for sharp bioinformatics skills, **Carsten Daub** – for allowing me to visit RIKEN. **Morana. Ellen Sherwood** – for answering questions during my first year. **Öjar Melefors, Alma Brolund, Linus Sandegren** and **Karin Tegmark Wisell** for interesting discussions about plasmid research. **Karin Troell** and **Johan Ankarklev**, for scientific discussions. **Rickard Sandberg** – for doing great research and allowing me to use the cluster. **Elin Einarsson** for 3' RACE. **Bengt Persson** for runtime at LiU. **Elsie Castro** for running 454. **Nicolas** and **Emilie** for friendship. **Carlos Talavera-López** for PCRs. My co-authors at LSHTM and UEA: Prof. **Michael Miles, Claire Butler, Louisa Messenger, Michael Lewis, Martin Llewellyn** and **Kevin Tyler**. **Anna Wetterbom** for chats and advice. **Hamid Darban, Stefanie Prast-Nielsen, Johannes Luthman** for chats about science and life. **Jan Andersson** and **Feifei Xu** for evolutionary insights. **Fredrik Lysholm** for technical chats.

---

**Daniel Nilsson** for answering questions, especially during my first year. The fellow PhD students I met on the comparative genomics course 2011 at Pasteur Institute. And to anyone I forgot: I apologise, it is entirely my fault.

I also acknowledge sources of funding; **KID-medel**, **FORMAS**, **Vetenskapsrådet**, as well as the people that were not directly involved in this piece of research: administrators, technical staff, annotators, anonymous reviewers, open source developers and all others that work to create a good research environment.

I acknowledge friends and family outside of research. **Robert** and **Christofer** for long-time friendship. **Magnus**, **Henrik** and their spouses for always inviting me. My sister and my parents: **Fanny**, **Peringe** and **Sinikka** for support.

Now that I reached the last lines I would like to wish past, present and future friends all the best of luck.

Oscar, October 3, 2012. Stockholm.



This thesis was typeset with L<sup>A</sup>T<sub>E</sub>X.



---

## 8 References

*Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.*

– M.S. Curie (1867-1934), physicist and chemist.

- [1] T. J. Anderson and J. Jaenike. “Host specificity, evolutionary relationships and macrogeographic differentiation among *Ascaris* populations from humans and pigs”. *Parasitology* 115 ( Pt 3) (1997), pp. 325–42.
- [2] T. H. Le, D. Blair, and D. P. McManus. “Mitochondrial genomes of human helminths and their use as markers in population genetics and phylogeny”. *Acta Trop* 77.3 (2000), pp. 243–56.
- [3] E. Pays, B. Vanhollebeke, L. Vanhamme, F. Paturiaux-Hanocq, D. P. Nolan, and D. Perez-Morga. “The trypanolytic factor of human serum”. *Nat Rev Microbiol* 4.6 (2006), pp. 477–86.
- [4] F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen. “Nucleotide sequence of bacteriophage lambda DNA”. *J Mol Biol* 162.4 (1982), pp. 729–73.
- [5] C. J. Metcalfe, J. Filee, I. Germon, J. Joss, and D. Casane. “Evolution of the Australian Lungfish (*Neoceratodus forsteri*) Genome: A Major Role for CR1 and L2 LINE Elements”. *Mol Biol Evol* (2012).
- [6] C. Alkan, S. Sajjadian, and E. E. Eichler. “Limitations of next-generation genome sequence assembly”. *Nat Methods* 8.1 (2011), pp. 61–5.
- [7] M. Wanunu. “Nanopores: A journey towards DNA sequencing”. *Phys Life Rev* 9.2 (2012), pp. 125–58.
- [8] C. Aurrecoechea, J. Brestelli, B. P. Brunk, S. Fischer, B. Gajria, et al. “EuPathDB: a portal to eukaryotic pathogen databases”. *Nucleic Acids Res* 38.Database issue (2010), pp. D415–9.
- [9] A. L. Price, N. C. Jones, and P. A. Pevzner. “De novo identification of repeat families in large genomes”. *Bioinformatics* 21 Suppl 1 (2005), pp. i351–8.
- [10] M. Tarailo-Graovac and N. Chen. “Using RepeatMasker to identify repetitive elements in genomic sequences”. *Curr Protoc Bioinformatics* Chapter 4 (2009), Unit 4 10.
- [11] Z. Wang, M. Gerstein, and M. Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. *Nat Rev Genet* 10.1 (2009), pp. 57–63.
- [12] X. Fu, N. Fu, S. Guo, Z. Yan, Y. Xu, et al. “Estimating accuracy of RNA-Seq and microarrays with proteomics”. *BMC Genomics* 10 (2009), p. 161.
- [13] M. D. Robinson and A. Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data”. *Genome Biol* 11.3 (2010), R25.

- 
- [14] J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, et al. "Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data". *Bioinformatics* 25.24 (2009), pp. 3207–12.
- [15] P. J. Shepard, E. A. Choi, J. Lu, L. A. Flanagan, K. J. Hertel, and Y. Shi. "Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq". *RNA* 17.4 (2011), pp. 761–72.
- [16] A. G. Simpson. "Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota)". *Int J Syst Evol Microbiol* 53.Pt 6 (2003), pp. 1759–77.
- [17] C. Dobell. "The Discovery of the Intestinal Protozoa of Man". *Proc R Soc Med* 13.Sect Hist Med (1920), pp. 1–15.
- [18] W. Lambl. "Mikroskopische untersuchungen der Darmexcrete [Article in German]". *Vierteljahrsschr. Prakt. Heikunde* 61 (1859), pp. 1–58.
- [19] R. D. Adam. "Biology of Giardia lamblia". *Clin Microbiol Rev* 14.3 (2001), pp. 447–75.
- [20] J. Ankarklev, J. Jerlstrom-Hultqvist, E. Ringqvist, K. Troell, and S. G. Svard. "Behind the smile: cell biology and disease mechanisms of Giardia species". *Nat Rev Microbiol* 8.6 (2010), pp. 413–22.
- [21] G. P. Kent, J. R. Greenspan, J. L. Herndon, L. M. Mofenson, J. A. Harris, et al. "Epidemic giardiasis caused by a contaminated public water supply". *Am J Public Health* 78.2 (1988), pp. 139–43.
- [22] R. E. Black, A. C. Dykes, S. P. Sinclair, and J. G. Wells. "Giardiasis in day-care centers: evidence of person-to-person transmission". *Pediatrics* 60.4 (1977), pp. 486–91.
- [23] K. Nygard, B. Schimmer, O. Sobstad, A. Walde, I. Tveit, et al. "A large community outbreak of waterborne giardiasis-delayed detection in a non-endemic urban area". *BMC Public Health* 6 (2006), p. 141.
- [24] L. J. Robertson, K. Hanevik, A. A. Escobedo, K. Morch, and N. Langeland. "Giardiasis—why do the symptoms sometimes never stop?" *Trends Parasitol* 26.2 (2010), pp. 75–82.
- [25] K. A. Wensaas, N. Langeland, K. Hanevik, K. Morch, G. E. Eide, and G. Rortveit. "Irritable bowel syndrome and chronic fatigue 3 years after acute giardiasis: historic cohort study". *Gut* 61.2 (2012), pp. 214–9.
- [26] L. Savioli, H. Smith, and A. Thompson. "Giardia and Cryptosporidium join the 'Neglected Diseases Initiative'". *Trends Parasitol* 22.5 (2006), pp. 203–8.
- [27] S. C. Lenaghan, C. A. Davis, W. R. Henson, Z. Zhang, and M. Zhang. "High-speed microscopic imaging of flagella motility and swimming in Giardia lamblia trophozoites". *Proc Natl Acad Sci U S A* 108.34 (2011), E550–8.
- [28] J. Hernandez-Sanchez, R. F. Linan, R. Salinas-Tobon Mdel, and G. Ortega-Pierres. "Giardia duodenalis: adhesion-deficient clones have reduced ability to establish infection in Mongolian gerbils". *Exp Parasitol* 119.3 (2008), pp. 364–72.
- [29] R. Bernander, J. E. Palm, and S. G. Svard. "Genome ploidy in different stages of the Giardia lamblia life cycle". *Cell Microbiol* 3.1 (2001), pp. 55–62.

- 
- [30] M. Abodeely, K. N. DuBois, A. Hehl, S. Stefanic, M. Sajid, et al. "A contiguous compartment functions as endoplasmic reticulum and endosome/lysosome in *Giardia lamblia*". *Eukaryot Cell* 8.11 (2009), pp. 1665–76.
- [31] J. Tovar, G. Leon-Avila, L. B. Sanchez, R. Sutak, J. Tachezy, et al. "Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation". *Nature* 426.6963 (2003), pp. 172–6.
- [32] R. C. Rendtorff. "The experimental transmission of human intestinal protozoan parasites. II. *Giardia lamblia* cysts given in capsules". *Am J Hyg* 59.2 (1954), pp. 209–20.
- [33] H. Troeger, H. J. Eppe, T. Schneider, U. Wahnschaffe, R. Ullrich, et al. "Effect of chronic *Giardia lamblia* infection on epithelial transport and barrier function in human duodenum". *Gut* 56.3 (2007), pp. 328–35.
- [34] T. B. de Carvalho, E. B. David, S. T. Coradi, and S. Guimaraes. "Protease activity in extracellular products secreted in vitro by trophozoites of *Giardia duodenalis*". *Parasitol Res* 104.1 (2008), pp. 185–90.
- [35] M. L. Sogin, J. H. Gunderson, H. J. Elwood, R. A. Alonso, and D. A. Peattie. "Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*". *Science* 243.4887 (1989), pp. 75–7.
- [36] D. D. Leipe, J. H. Gunderson, T. A. Nerad, and M. L. Sogin. "Small subunit ribosomal RNA+ of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree". *Mol Biochem Parasitol* 59.1 (1993), pp. 41–8.
- [37] T. Hashimoto, Y. Nakamura, T. Kamaishi, F. Nakamura, J. Adachi, et al. "Phylogenetic place of mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2". *Mol Biol Evol* 12.5 (1995), pp. 782–93.
- [38] E. Hilario and J. P. Gogarten. "The prokaryote-to-eukaryote transition reflected in the evolution of the V/F/A-ATPase catalytic and proteolipid subunits". *J Mol Evol* 46.6 (1998), pp. 703–15.
- [39] A. J. Roger, S. G. Svard, J. Tovar, C. G. Clark, M. W. Smith, et al. "A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria". *Proc Natl Acad Sci U S A* 95.1 (1998), pp. 229–34.
- [40] T. Hashimoto, L. B. Sanchez, T. Shirakura, M. Muller, and M. Hasegawa. "Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny". *Proc Natl Acad Sci U S A* 95.12 (1998), pp. 6860–5.
- [41] L. F. Jimenez-Garcia, G. Zavala, B. Chavez-Munguia, P. Ramos-Godinez Mdel, G. Lopez-Velazquez, et al. "Identification of nucleoli in the early branching protist *Giardia duodenalis*". *Int J Parasitol* 38.11 (2008), pp. 1297–304.
- [42] J. B. Dacks, A. Marinets, W. Ford Doolittle, T. Cavalier-Smith, and Jr. Logsdon J. M. "Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang". *Mol Biol Evol* 19.6 (2002), pp. 830–40.



- 
- [43] H. Brinkmann, M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. "An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics". *Syst Biol* 54.5 (2005), pp. 743–57.
- [44] J. Luo, M. Teng, G. P. Zhang, Z. R. Lun, H. Zhou, and L. H. Qu. "Evaluating the evolution of *G. lamblia* based on the small nucleolar RNAs identified from Archaea and unicellular eukaryotes". *Parasitol Res* 104.6 (2009), pp. 1543–6.
- [45] M. Tibayrenc, F. Kjellberg, and F. J. Ayala. "A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences". *Proc Natl Acad Sci U S A* 87.7 (1990), pp. 2414–8.
- [46] M. A. Ramesh, S. B. Malik, and Jr. Logsdon J. M. "A phylogenomic inventory of meiotic genes; evidence for sex in *Giardia* and an early eukaryotic origin of meiosis". *Curr Biol* 15.2 (2005), pp. 185–91.
- [47] E. Lasek-Nesselquist, D. M. Welch, R. C. Thompson, R. F. Steuart, and M. L. Sogin. "Genetic exchange within and between assemblages of *Giardia duodenalis*". *J Eukaryot Microbiol* 56.6 (2009), pp. 504–18.
- [48] M. W. Gaunt, M. Yeo, I. A. Frame, J. R. Stothard, H. J. Carrasco, et al. "Mechanism of genetic exchange in American trypanosomes". *Nature* 421.6926 (2003), pp. 936–9.
- [49] M. W. Black and J. C. Boothroyd. "Lytic cycle of *Toxoplasma gondii*". *Microbiol Mol Biol Rev* 64.3 (2000), pp. 607–23.
- [50] M. C. Bruce, P. Alano, S. Duthie, and R. Carter. "Commitment of the malaria parasite *Plasmodium falciparum* to sexual and asexual development". *Parasitology* 100 Pt 2 (1990), pp. 191–200.
- [51] Y. Feng and L. Xiao. "Zoonotic potential and molecular epidemiology of *Giardia* species and giardiasis". *Clin Microbiol Rev* 24.1 (2011), pp. 110–40.
- [52] A. Thompson. "Human giardiasis: genotype-linked differences in clinical symptomatology". *Trends Parasitol* 17.10 (2001), p. 465.
- [53] J. Sahagun, A. Clavel, P. Goni, C. Seral, M. T. Llorente, et al. "Correlation between the presence of symptoms and the *Giardia duodenalis* genotype". *Eur J Clin Microbiol Infect Dis* 27.1 (2008), pp. 81–3.
- [54] T. E. Nash, D. A. Herrington, G. A. Losonsky, and M. M. Levine. "Experimental human infections with *Giardia lamblia*". *J Infect Dis* 156.6 (1987), pp. 974–84.
- [55] E. Lasek-Nesselquist, D. M. Welch, and M. L. Sogin. "The identification of a new *Giardia duodenalis* assemblage in marine vertebrates and a preliminary analysis of *G. duodenalis* population biology in marine systems". *Int J Parasitol* 40.9 (2010), pp. 1063–74.
- [56] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods". *Mol Biol Evol* 28.10 (2011), pp. 2731–9.

- 
- [57] S. L. Erlandsen, W. J. Bemrick, C. L. Wells, D. E. Feely, L. Knudson, et al. "Axenic culture and characterization of *Giardia ardeae* from the great blue heron (*Ardea herodias*)". *J Parasitol* 76.5 (1990), pp. 717–24.
- [58] P. T. Monis, S. M. Caccio, and R. C. Thompson. "Variation in *Giardia*: towards a taxonomic revision of the genus". *Trends Parasitol* 25.2 (2009), pp. 93–100.
- [59] J. A. Upcroft, P. F. Boreham, and P. Upcroft. "Geographic variation in *Giardia* karyotypes". *Int J Parasitol* 19.5 (1989), pp. 519–27.
- [60] K. S. Kabnick and D. A. Peattie. "In situ analyses reveal that the two nuclei of *Giardia lamblia* are equivalent". *J Cell Sci* 95 ( Pt 3) (1990), pp. 353–60.
- [61] L. Z. Yu, Jr. Birky C. W., and R. D. Adam. "The two nuclei of *Giardia* each have complete copies of the genome and are partitioned equationally at cytokinesis". *Eukaryot Cell* 1.2 (2002), pp. 191–9.
- [62] H. G. Morrison, A. G. McArthur, F. D. Gillin, S. B. Aley, R. D. Adam, et al. "Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*". *Science* 317.5846 (2007), pp. 1921–6.
- [63] J. E. Nixon, A. Wang, H. G. Morrison, A. G. McArthur, M. L. Sogin, et al. "A spliceosomal intron in *Giardia lamblia*". *Proc Natl Acad Sci U S A* 99.6 (2002), pp. 3701–5.
- [64] A. G. Russell, T. E. Shutt, R. F. Watkins, and M. W. Gray. "An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of *Giardia lamblia*". *BMC Evol Biol* 5 (2005), p. 45.
- [65] S. W. Roy, A. J. Hudson, J. Joseph, J. Yee, and A. G. Russell. "Numerous fragmented spliceosomal introns, AT-AC splicing, and an unusual dynein gene expression pathway in *Giardia lamblia*". *Mol Biol Evol* 29.1 (2012), pp. 43–9.
- [66] R. Kamikawa, Y. Inagaki, M. Tokoro, A. J. Roger, and T. Hashimoto. "Split introns in the genome of *Giardia intestinalis* are excised by spliceosome-mediated trans-splicing". *Curr Biol* 21.4 (2011), pp. 311–5.
- [67] G. Manning, D. S. Reiner, T. Lauwaet, M. Dacre, A. Smith, et al. "The minimal kinome of *Giardia lamblia* illuminates early kinase evolution and unique parasite biology". *Genome Biol* 12.7 (2011), R66.
- [68] R. D. Adam, A. Nigam, V. Seshadri, C. A. Martens, G. A. Farneth, et al. "The *Giardia lamblia* vsp gene repertoire: characteristics, genomic organization, and evolution". *BMC Genomics* 11 (2010), p. 424.
- [69] T. E. Nash, H. T. Lujan, M. R. Mowatt, and J. T. Conrad. "Variant-specific surface protein switching in *Giardia lamblia*". *Infect Immun* 69.3 (2001), pp. 1922–3.
- [70] T. E. Nash, S. M. Banks, D. W. Alling, Jr. Merritt J. W., and J. T. Conrad. "Frequency of variant antigens in *Giardia lamblia*". *Exp Parasitol* 71.4 (1990), pp. 415–21.
- [71] L. Kulakova, S. M. Singer, J. Conrad, and T. E. Nash. "Epigenetic mechanisms are involved in the control of *Giardia lamblia* antigenic variation". *Mol Microbiol* 61.6 (2006), pp. 1533–42.

- 
- [72] C. G. Prucca, I. Slavin, R. Quiroga, E. V. Elias, F. D. Rivero, et al. "Antigenic variation in *Giardia lamblia* is regulated by RNA interference". *Nature* 456.7223 (2008), pp. 750–4.
- [73] I. R. Arkhipova and H. G. Morrison. "Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead". *Proc Natl Acad Sci U S A* 98.25 (2001), pp. 14497–502.
- [74] E. Ullu, H. D. Lujan, and C. Tschudi. "Small sense and antisense RNAs derived from a telomeric retroposon family in *Giardia intestinalis*". *Eukaryot Cell* 4.6 (2005), pp. 1155–7.
- [75] D. Mark Welch and M. Meselson. "Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange". *Science* 288.5469 (2000), pp. 1211–5.
- [76] A. R. Omilian, M. E. Cristescu, J. L. Dudycha, and M. Lynch. "Ameiotic recombination in asexual lineages of *Daphnia*". *Proc Natl Acad Sci U S A* 103.49 (2006), pp. 18638–43.
- [77] M. K. Poxleitner, M. L. Carpenter, J. J. Mancuso, C. J. Wang, S. C. Dawson, and W. Z. Cande. "Evidence for karyogamy and exchange of genetic material in the binucleate intestinal parasite *Giardia intestinalis*". *Science* 319.5869 (2008), pp. 1530–3.
- [78] M. L. Carpenter, Z. J. Assaf, S. Gourguechon, and W. Z. Cande. "Nuclear inheritance and genetic exchange without meiosis in the binucleate parasite *Giardia intestinalis*". *J Cell Sci* 125.Pt 10 (2012), pp. 2523–32.
- [79] M. Benchimol. "*Giardia lamblia*: behavior of the nuclear envelope". *Parasitol Res* 94.4 (2004), pp. 254–264.
- [80] P. Tumova, K. Hofstetrova, E. Nohynkova, O. Hovorka, and J. Kral. "Cytogenetic evidence for diversity of two nuclei within a single diplomonad cell of *Giardia*". *Chromosoma* 116.1 (2007), pp. 65–78.
- [81] A. A. Saraiya, W. Li, and C. C. Wang. "A microRNA derived from an apparent canonical biogenesis pathway regulates variant surface protein gene expression in *Giardia lamblia*". *RNA* 17.12 (2011), pp. 2152–64.
- [82] Y. Yang and R. D. Adam. "Allele-specific expression of a variant-specific surface protein (VSP) of *Giardia lamblia*". *Nucleic Acids Res* 22.11 (1994), pp. 2102–8.
- [83] A. A. Best, H. G. Morrison, A. G. McArthur, M. L. Sogin, and G. J. Olsen. "Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*". *Genome Res* 14.8 (2004), pp. 1537–47.
- [84] S. R. Birkeland, S. P. Preheim, B. J. Davids, M. J. Cipriano, D. Palm, et al. "Transcriptome analyses of the *Giardia lamblia* life cycle". *Mol Biochem Parasitol* 174.1 (2010), pp. 62–5.
- [85] L. Morf, C. Spycher, H. Rehrauer, C. A. Fournier, H. G. Morrison, and A. B. Hehl. "The transcriptional response to encystation stimuli in *Giardia lamblia* is restricted to a small set of genes". *Eukaryot Cell* 9.10 (2010), pp. 1566–76.

- 
- [86] L. A. Knodler, S. G. Svard, J. D. Silberman, B. J. Davids, and F. D. Gillin. "Developmental gene regulation in *Giardia lamblia*: first evidence for an encystation-specific promoter and differential 5' mRNA processing". *Mol Microbiol* 34.2 (1999), pp. 327–40.
- [87] C. H. Sun, D. Palm, A. G. McArthur, S. G. Svard, and F. D. Gillin. "A novel Myb-related protein involved in transcriptional activation of encystation genes in *Giardia lamblia*". *Mol Microbiol* 46.4 (2002), pp. 971–84.
- [88] J. Yee and P. P. Dennis. "Isolation and characterization of a NADP-dependent glutamate dehydrogenase gene from the primitive eucaryote *Giardia lamblia*". *J Biol Chem* 267.11 (1992), pp. 7539–44.
- [89] D. V. Holberton and J. Marshall. "Analysis of consensus sequence patterns in *Giardia* cytoskeleton gene promoters". *Nucleic Acids Res* 23.15 (1995), pp. 2945–53.
- [90] H. G. Elmendorf, S. M. Singer, J. Pierce, J. Cowan, and T. E. Nash. "Initiator and upstream elements in the alpha2-tubulin promoter of *Giardia lamblia*". *Mol Biochem Parasitol* 113.1 (2001), pp. 157–69.
- [91] S. Teodorovic, C. D. Walls, and H. G. Elmendorf. "Bidirectional transcription is an inherent feature of *Giardia lamblia* promoters and contributes to an abundance of sterile antisense transcripts throughout the genome". *Nucleic Acids Res* 35.8 (2007), pp. 2544–53.
- [92] A. A. Saraiya and C. C. Wang. "snoRNA, a novel precursor of microRNA in *Giardia lamblia*". *PLoS Pathog* 4.11 (2008), e1000224.
- [93] I. J. Macrae, K. Zhou, F. Li, A. Repic, A. N. Brooks, et al. "Structural basis for double-stranded RNA processing by Dicer". *Science* 311.5758 (2006), pp. 195–8.
- [94] A. L. Wang, H. M. Yang, K. A. Shen, and C. C. Wang. "*Giardiavirus* double-stranded RNA genome encodes a capsid polypeptide and a gag-pol-like fusion protein by a translation frameshift". *Proc Natl Acad Sci U S A* 90.18 (1993), pp. 8595–9.
- [95] C. Chagas. "Nova tripanozomíase humana: estudos sobre a morfologia e o ciclo evolutivo do *Schizotrypanum cruzi* n. gen., n. sp., agente etiológico de nova entidade morbida do homem [Article in Portuguese]". pt. *Mem. Inst. Oswaldo Cruz* 1 (Aug. 1909), pp. 159–218.
- [96] R. L. Tarleton, R. Reithinger, J. A. Urbina, U. Kitron, and R. E. Gurtler. "The challenges of Chagas Disease—grim outlook or glimmer of hope". *PLoS Med* 4.12 (2007), e332.
- [97] Jr. Rassi A., A. Rassi, and J. A. Marin-Neto. "Chagas disease". *Lancet* 375.9723 (2010), pp. 1388–402.
- [98] G. A. Schmunis. "Epidemiology of Chagas disease in non-endemic countries: the role of international migration". *Mem Inst Oswaldo Cruz* 102 Suppl 1 (2007), pp. 75–85.
- [99] M. A. Miles, M. S. Llewellyn, M. D. Lewis, M. Yeo, R. Baleela, et al. "The molecular epidemiology and phylogeography of *Trypanosoma cruzi* and parallel research on *Leishmania*: looking back and to the future". *Parasitology* 136.12 (2009), pp. 1509–28.

- 
- [100] A. C. Aufderheide, W. Salo, M. Madden, J. Streitz, J. Buikstra, et al. "A 9,000-year record of Chagas' disease". *Proc Natl Acad Sci U S A* 101.7 (2004), pp. 2034–9.
- [101] M. J. Pinazo, G. Espinosa, M. Gallego, P. L. Lopez-Chejade, J. A. Urbina, and J. Gascon. "Successful treatment with posaconazole of a patient with chronic Chagas disease and systemic lupus erythematosus". *Am J Trop Med Hyg* 82.4 (2010), pp. 583–7.
- [102] L. Simpson and J. Shaw. "RNA editing and the mitochondrial cryptogenes of kinetoplastid protozoa". *Cell* 57.3 (1989), pp. 355–66.
- [103] D. A. Campbell, S. Thomas, and N. R. Sturm. "Transcription in kinetoplastid protozoa: why be normal?" *Microbes Infect* 5.13 (2003), pp. 1231–40.
- [104] J. D. Barry and R. McCulloch. "Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite". *Adv Parasitol* 49 (2001), pp. 1–70.
- [105] C. De Geer. "Mémoires pour servir à l'histoire des insectes. [Publication in French]". *Grefing and Hesselberg. (Stockholm)* (1773).
- [106] C. Young, P. Losikoff, A. Chawla, L. Glasser, and E. Forman. "Transfusion-acquired *Trypanosoma cruzi* infection". *Transfusion* 47.3 (2007), pp. 540–4.
- [107] Control Centers for Disease and Prevention. "Chagas disease after organ transplantation—United States, 2001". *MMWR Morb Mortal Wkly Rep* 51.10 (2002), pp. 210–2.
- [108] R. E. Gurtler, E. L. Segura, and J. E. Cohen. "Congenital transmission of *Trypanosoma cruzi* infection in Argentina". *Emerg Infect Dis* 9.1 (2003), pp. 29–32.
- [109] P. L. Dorn, L. Perniciaro, M. J. Yabsley, D. M. Roellig, G. Balsamo, et al. "Autochthonous transmission of *Trypanosoma cruzi*, Louisiana". *Emerg Infect Dis* 13.4 (2007), pp. 605–7.
- [110] A. A. Nobrega, M. H. Garcia, E. Tatto, M. T. Obara, E. Costa, et al. "Oral transmission of Chagas disease by consumption of acai palm fruit, Brazil". *Emerg Infect Dis* 15.4 (2009), pp. 653–5.
- [111] C. J. Bastos, R. Aras, G. Mota, F. Reis, J. P. Dias, et al. "Clinical outcomes of thirteen patients with acute chagas disease acquired through oral transmission from two urban outbreaks in northeastern Brazil". *PLoS Negl Trop Dis* 4.6 (2010), e711.
- [112] M. A. Shikanai-Yasuda and N. B. Carvalho. "Oral transmission of Chagas disease". *Clin Infect Dis* 54.6 (2012), pp. 845–52.
- [113] J. M. Hofflin, R. H. Sadler, F. G. Araujo, W. E. Page, and J. S. Remington. "Laboratory-acquired Chagas disease". *Trans R Soc Trop Med Hyg* 81.3 (1987), pp. 437–40.
- [114] J. C. Dias, A. C. Silveira, and C. J. Schofield. "The impact of Chagas disease control in Latin America: a review". *Mem Inst Oswaldo Cruz* 97.5 (2002), pp. 603–12.
- [115] R. E. Bernstein. "Darwin's illness: Chagas' disease resurges". *J R Soc Med* 77.7 (1984), pp. 608–9.
- [116] K. M. Tyler and D. M. Engman. "The life cycle of *Trypanosoma cruzi* revisited". *Int J Parasitol* 31.5-6 (2001), pp. 472–81.

- 
- [117] M. C. Bonaldo, T. Souto-Padron, W. de Souza, and S. Goldenberg. "Cell-substrate adhesion during *Trypanosoma cruzi* differentiation". *J Cell Biol* 106.4 (1988), pp. 1349–58.
  - [118] I. Tardieux, P. Webster, J. Ravesloot, W. Boron, J. A. Lunn, et al. "Lysosome recruitment and fusion are early events required for trypanosome invasion of mammalian cells". *Cell* 71.7 (1992), pp. 1117–30.
  - [119] A. Rodriguez, E. Samoff, M. G. Rioult, A. Chung, and N. W. Andrews. "Host cell invasion by trypanosomes requires lysosomes and microtubule/kinesin-mediated transport". *J Cell Biol* 134.2 (1996), pp. 349–62.
  - [120] N. W. Andrews. "Living dangerously: how *Trypanosoma cruzi* uses lysosomes to get inside host cells, and then escapes into the cytoplasm". *Biol Res* 26.1-2 (1993), pp. 65–7.
  - [121] C. A. Hoare. "The Trypanosomes of Mammals". *Blackwell Scientific Publications, Oxford* (1972).
  - [122] J. R. Baker, M. A. Miles, D. G. Godfrey, and T. V. Barrett. "Biochemical characterization of some species of *Trypanosoma* (Schizotrypanum) from bats (Microchiroptera)". *Am J Trop Med Hyg* 27.3 (1978), pp. 483–91.
  - [123] M. D. Lewis, M. S. Llewellyn, M. Yeo, N. Acosta, M. W. Gaunt, and M. A. Miles. "Recent, independent and anthropogenic origins of *Trypanosoma cruzi* hybrids". *PLoS Negl Trop Dis* 5.10 (2011), e1363.
  - [124] S. Brisse, J. Henriksson, C. Barnabe, E. J. Douzery, D. Berkvens, et al. "Evidence for genetic exchange and hybridization in *Trypanosoma cruzi* based on nucleotide sequences and molecular karyotype". *Infect Genet Evol* 2.3 (2003), pp. 173–83.
  - [125] B. Zingales, M. A. Miles, D. A. Campbell, M. Tibayrenc, A. M. Macedo, et al. "The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications". *Infect Genet Evol* 12.2 (2012), pp. 240–53.
  - [126] A. Marcili, L. Lima, M. Cavazzana, A. C. Junqueira, H. H. Veludo, et al. "A new genotype of *Trypanosoma cruzi* associated with bats evidenced by phylogenetic analyses using SSU rDNA, cytochrome b and Histone H2B genes and genotyping based on ITS1 rDNA". *Parasitology* 136.6 (2009), pp. 641–55.
  - [127] J. A. Lake, V. F. de la Cruz, P. C. Ferreira, C. Morel, and L. Simpson. "Evolution of parasitism: kinetoplastid protozoan history reconstructed from mitochondrial rRNA gene sequences". *Proc Natl Acad Sci U S A* 85.13 (1988), pp. 4779–83.
  - [128] A. P. Fernandes, K. Nelson, and S. M. Beverley. "Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of parasitism". *Proc Natl Acad Sci U S A* 90.24 (1993), pp. 11608–12.
  - [129] D. A. Maslov, J. Lukes, M. Jirku, and L. Simpson. "Phylogeny of trypanosomes as inferred from the small and large subunit rRNAs: implications for the evolution of parasitism in the trypanosomatid protozoa". *Mol Biochem Parasitol* 75.2 (1996), pp. 197–205.
  - [130] J. R. Stevens and W. Gibson. "The molecular evolution of trypanosomes". *Parasitol Today* 15.11 (1999), pp. 432–7.

- 
- [131] S. L. Bonatto and F. M. Salzano. "A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data". *Proc Natl Acad Sci U S A* 94.5 (1997), pp. 1866–71.
- [132] M. E. Thomas, J. J. Rasweiler Iv, and A. D'Alessandro. "Experimental transmission of the parasitic flagellates *Trypanosoma cruzi* and *Trypanosoma rangeli* between triatomine bugs or mice and captive neotropical bats". *Mem Inst Oswaldo Cruz* 102.5 (2007), pp. 559–65.
- [133] P. B. Hamilton, M. M. Teixeira, and J. R. Stevens. "The evolution of *Trypanosoma cruzi*: the 'bat seeding' hypothesis". *Trends Parasitol* 28.4 (2012), pp. 136–41.
- [134] L. Lima, F. M. Silva, L. Neves, M. Attias, C. S. Takata, et al. "Evolutionary Insights from Bat Trypanosomes: Morphological, Developmental and Phylogenetic Evidence of a New Species, *Trypanosoma* (Schizotrypanum) *erneyi* sp. nov., in African Bats Closely Related to *Trypanosoma* (Schizotrypanum) *cruzi* and Allied Species". *Protist* (2012).
- [135] C. V. Lisboa, A. P. Pinho, H. M. Herrera, M. Gerhardt, E. Cupolillo, and A. M. Jansen. "*Trypanosoma cruzi* (Kinetoplastida, Trypanosomatidae) genotypes in neotropical bats in Brazil". *Vet Parasitol* 156.3-4 (2008), pp. 314–8.
- [136] F. Maia da Silva, A. Marcili, L. Lima, Jr. Cavazzana M., P. A. Ortiz, et al. "*Trypanosoma rangeli* isolates of bats from Central Brazil: genotyping and phylogenetic analysis enable description of a new lineage using spliced-leader gene sequences". *Acta Trop* 109.3 (2009), pp. 199–207.
- [137] Jr. Cavazzana M., A. Marcili, L. Lima, F. M. da Silva, A. C. Junqueira, et al. "Phylogeographical, ecological and biological patterns shown by nuclear (ssrRNA and gGAPDH) and mitochondrial (Cyt b) genes of trypanosomes of the subgenus *Schizotrypanum* parasitic in Brazilian bats". *Int J Parasitol* 40.3 (2010), pp. 345–55.
- [138] P. B. Hamilton, C. Cruickshank, J. R. Stevens, M. M. Teixeira, and F. Mathews. "Parasites reveal movement of bats between the New and Old Worlds". *Mol Phylogenet Evol* 63.2 (2012), pp. 521–6.
- [139] D. R. Brooks and A. L. Ferraio. "The historical biogeography of co-evolution: emerging infectious diseases are evolutionary accidents waiting to happen". *Journal of Biogeography* 32.8 (2005), pp. 1291–1299.
- [140] P. B. Hamilton, W. C. Gibson, and J. R. Stevens. "Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies". *Mol Phylogenet Evol* 44.1 (2007), pp. 15–25.
- [141] D. M. Engman, L. V. Reddy, J. E. Donelson, and L. V. Kirchhoff. "*Trypanosoma cruzi* exhibits inter- and intra-strain heterogeneity in molecular karyotype and chromosomal gene location". *Mol Biochem Parasitol* 22.2-3 (1987), pp. 115–23.
- [142] W. Wagner and M. So. "Genomic variation of *Trypanosoma cruzi*: involvement of multicopy genes". *Infect Immun* 58.10 (1990), pp. 3217–24.
- [143] J. Henriksson, B. Porcel, M. Rydaker, A. Ruiz, V. Sabaj, et al. "Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi*". *Mol Biochem Parasitol* 73.1-2 (1995), pp. 63–74.

- 
- [144] M. D. Lewis, M. S. Llewellyn, M. W. Gaunt, M. Yeo, H. J. Carrasco, and M. A. Miles. "Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanosoma cruzi* populations and expose contrasts between natural and experimental hybrids". *Int J Parasitol* 39.12 (2009), pp. 1305–17.
- [145] N. M. El-Sayed, P. J. Myler, D. C. Bartholomeu, D. Nilsson, G. Aggarwal, et al. "The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease". *Science* 309.5733 (2005), pp. 409–15.
- [146] B. Zingales, M. E. Pereira, K. A. Almeida, E. S. Umezawa, N. S. Nehme, et al. "Biological parameters and molecular markers of clone CL Brener—the reference organism of the *Trypanosoma cruzi* genome project". *Mem Inst Oswaldo Cruz* 92.6 (1997), pp. 811–4.
- [147] Z. Brener and E. Chiari. "Morphological Variations Observed in Different Strains of *Trypanosoma Cruzi* [Article in Portuguese]". *Rev Inst Med Trop Sao Paulo* 5 (1963), pp. 220–4.
- [148] M. Berriman, E. Ghedin, C. Hertz-Fowler, G. Blandin, H. Renauld, et al. "The genome of the African trypanosome *Trypanosoma brucei*". *Science* 309.5733 (2005), pp. 416–22.
- [149] A. C. Ivens, C. S. Peacock, E. A. Wortley, L. Murphy, G. Aggarwal, et al. "The genome of the kinetoplastid parasite, *Leishmania major*". *Science* 309.5733 (2005), pp. 436–42.
- [150] D. B. Weatherly, C. Boehlke, and R. L. Tarleton. "Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome". *BMC Genomics* 10 (2009), p. 255.
- [151] E. Arner, E. Kindlund, D. Nilsson, F. Farzana, M. Ferella, et al. "Database of *Trypanosoma cruzi* repeated genes: 20,000 additional gene variants". *BMC Genomics* 8 (2007), p. 391.
- [152] A. C. Frasch. "Functional diversity in the trans-sialidase and mucin families in *Trypanosoma cruzi*". *Parasitol Today* 16.7 (2000), pp. 282–6.
- [153] J. Neres, R. A. Bryce, and K. T. Douglas. "Rational drug design in parasitology: trans-sialidase as a case study for Chagas disease". *Drug Discov Today* 13.3–4 (2008), pp. 110–7.
- [154] T. A. Minning, D. B. Weatherly, S. Flibotte, and R. L. Tarleton. "Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in *Trypanosoma cruzi* strains revealed by array comparative genomic hybridization". *BMC Genomics* 12 (2011), p. 139.
- [155] P. Respuela, M. Ferella, A. Rada-Iglesias, and L. Aslund. "Histone acetylation and methylation at sites initiating divergent polycistronic transcription in *Trypanosoma cruzi*". *J Biol Chem* 283.23 (2008), pp. 15884–92.
- [156] D. K. Ekanayake, T. Minning, B. Weatherly, K. Gunasekera, D. Nilsson, et al. "Epigenetic regulation of transcription and virulence in *Trypanosoma cruzi* by O-linked thymine glucosylation of DNA". *Mol Cell Biol* 31.8 (2011), pp. 1690–700.



- 
- [157] D. Ekanayake and R. Sabatini. “Epigenetic regulation of polymerase II transcription initiation in *Trypanosoma cruzi*: modulation of nucleosome abundance, histone modification, and polymerase occupancy by O-linked thymine DNA glucosylation”. *Eukaryot Cell* 10.11 (2011), pp. 1465–72.
- [158] F. van Leeuwen, M. C. Taylor, A. Mondragon, H. Moreau, W. Gibson, et al. “beta-D-glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres”. *Proc Natl Acad Sci U S A* 95.5 (1998), pp. 2366–71.
- [159] D. Dooijes, I. Chaves, R. Kieft, A. Dirks-Mulder, W. Martin, and P. Borst. “Base J originally found in kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*”. *Nucleic Acids Res* 28.16 (2000), pp. 3017–21.
- [160] L. F. Landweber and W. Gilbert. “Phylogenetic analysis of RNA editing: a primitive genetic phenomenon”. *Proc Natl Acad Sci U S A* 91.3 (1994), pp. 918–21.
- [161] A. Simoes-Barbosa, E. R. Arganaraz, A. M. Barros, C. Rosa Ade, N. P. Alves, et al. “Hitchhiking *Trypanosoma cruzi* minicircle DNA affects gene expression in human host cells via LINE-1 retrotransposon”. *Mem Inst Oswaldo Cruz* 101.8 (2006), pp. 833–43.
- [162] M. M. Hecht, N. Nitz, P. F. Araujo, A. O. Sousa, C. Rosa Ade, et al. “Inheritance of DNA transferred from American trypanosomes to human hosts”. *PLoS One* 5.2 (2010), e9181.
- [163] C. A. Brosnan and O. Voinnet. “The long and the short of noncoding RNAs”. *Curr Opin Cell Biol* 21.3 (2009), pp. 416–25.
- [164] L. F. Lye, K. Owens, H. Shi, S. M. Murta, A. C. Vieira, et al. “Retention and loss of RNA interference pathways in trypanosomatid protozoans”. *PLoS Pathog* 6.10 (2010), e1001161.
- [165] H. Ngo, C. Tschudi, K. Gull, and E. Ullu. “Double-stranded RNA induces mRNA degradation in *Trypanosoma brucei*”. *Proc Natl Acad Sci U S A* 95.25 (1998), pp. 14687–92.
- [166] Z. Wang, J. C. Morris, M. E. Drew, and P. T. Englund. “Inhibition of *Trypanosoma brucei* gene expression by RNA interference using an integratable vector with opposing T7 promoters”. *J Biol Chem* 275.51 (2000), pp. 40174–9.
- [167] W. D. DaRocha, K. Otsu, S. M. Teixeira, and J. E. Donelson. “Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi*”. *Mol Biochem Parasitol* 133.2 (2004), pp. 175–86.
- [168] A. Djikeng, H. Shi, C. Tschudi, and E. Ullu. “RNA interference in *Trypanosoma brucei*: cloning of small interfering RNAs provides evidence for retroposon-derived 24-26-nucleotide RNAs”. *RNA* 7.11 (2001), pp. 1522–30.
- [169] Y. Z. Wen, L. L. Zheng, J. Y. Liao, M. H. Wang, Y. Wei, et al. “Pseudogene-derived small interference RNAs regulate gene expression in African *Trypanosoma brucei*”. *Proc Natl Acad Sci U S A* 108.20 (2011), pp. 8345–50.
- [170] S. Michaeli, T. Doniger, S. K. Gupta, O. Wurtzel, M. Romano, et al. “RNA-seq analysis of small RNPs in *Trypanosoma brucei* reveals a rich repertoire of non-coding RNAs”. *Nucleic Acids Res* 40.3 (2012), pp. 1282–98.

- 
- [171] E. Ullu, C. Tschudi, and T. Chakraborty. "RNA interference in protozoan parasites". *Cell Microbiol* 6.6 (2004), pp. 509–19.
- [172] R. Hernandez, G. Nava, and M. Castaneda. "Small-size ribosomal RNA species in *Trypanosoma cruzi*". *Mol Biochem Parasitol* 8.4 (1983), pp. 297–304.
- [173] M. Milhausen, R. G. Nelson, S. Sather, M. Selkirk, and N. Agabian. "Identification of a small RNA containing the trypanosome spliced leader: a donor of shared 5' sequences of trypanosomatid mRNAs?" *Cell* 38.3 (1984), pp. 721–9.
- [174] M. R. Garcia-Silva, M. Frugier, J. P. Tosar, A. Correa-Dominguez, L. Ronalte-Alves, et al. "A population of tRNA-derived small RNAs is actively produced in *Trypanosoma cruzi* and recruited to specific cytoplasmic granules". *Mol Biochem Parasitol* 171.2 (2010), pp. 64–73.
- [175] P. D. Smith, F. D. Gillin, W. M. Spira, and T. E. Nash. "Chronic giardiasis: studies on drug sensitivity, toxin production, and host immune response". *Gastroenterology* 83.4 (1982), pp. 797–803.
- [176] A. Aggarwal, Jr. Merritt J. W., and T. E. Nash. "Cysteine-rich variant surface proteins of *Giardia lamblia*". *Mol Biochem Parasitol* 32.1 (1989), pp. 39–47.
- [177] B. Koudela, E. Nohynkova, J. Vitovec, M. Pakandl, and J. Kulda. "Giardia infection in pigs: detection and in vitro isolation of trophozoites of the *Giardia intestinalis* group". *Parasitology* 102 Pt 2 (1991), pp. 163–6.
- [178] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, et al. "Genome sequencing in microfabricated high-density picolitre reactors". *Nature* 437.7057 (2005), pp. 376–80.
- [179] C. Aurrecochea, J. Brestelli, B. P. Brunk, J. M. Carlton, J. Dommer, et al. "GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*". *Nucleic Acids Res* 37.Database issue (2009), pp. D526–30.
- [180] L. K. Fritz-Laylin, S. E. Prochnik, M. L. Ginger, J. B. Dacks, M. L. Carpenter, et al. "The genome of *Naegleria gruberi* illuminates early eukaryotic versatility". *Cell* 140.5 (2010), pp. 631–42.
- [181] M. Tibayrenc and F. J. Ayala. "Reproductive clonality of pathogens: A perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa". *Proc Natl Acad Sci U S A* (2012).
- [182] M. Postan, J. A. Dvorak, and J. P. McDaniel. "Studies of *Trypanosoma cruzi* clones in inbred mice. I. A comparison of the course of infection of C3H/HEN- mice with two clones isolated from a common source". *Am J Trop Med Hyg* 32.3 (1983), pp. 497–506.
- [183] M. B. Rogers, J. D. Hilley, N. J. Dickens, J. Wilkes, P. A. Bates, et al. "Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*". *Genome Res* 21.12 (2011), pp. 2129–42.
- [184] A. Dereeper, S. Audic, J. M. Claverie, and G. Blanc. "BLAST-EXPLORER helps you building datasets for phylogenetic analysis". *BMC Evol Biol* 10 (2010), p. 8.

- 
- [185] G. Talavera and J. Castresana. "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments". *Syst Biol* 56.4 (2007), pp. 564–77.
- [186] A. Stamatakis, T. Ludwig, and H. Meier. "RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees". *Bioinformatics* 21.4 (2005), pp. 456–63.
- [187] T. M. Lowe and S. R. Eddy. "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence". *Nucleic Acids Res* 25.5 (1997), pp. 955–64.
- [188] K. Darty, A. Denise, and Y. Ponty. "VARNA: Interactive drawing and editing of the RNA secondary structure". *Bioinformatics* 25.15 (2009), pp. 1974–5.
- [189] J. Jerlstrom-Hultqvist, O. Franzen, J. Ankarklev, F. Xu, E. Nohynkova, et al. "Genome analysis and comparative genomics of a *Giardia intestinalis* assemblage E isolate". *BMC Genomics* 11 (2010), p. 543.
- [190] O. Franzen, J. Jerlstrom-Hultqvist, E. Castro, E. Sherwood, J. Ankarklev, et al. "Draft genome sequencing of giardia intestinalis assemblage B isolate GS: is human giardiasis caused by two different species?" *PLoS Pathog* 5.8 (2009), e1000560.
- [191] H. Chen and P. C. Boutros. "VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R". *BMC Bioinformatics* 12 (2011), p. 35.
- [192] M. Kimura. "Evolutionary rate at the molecular level". *Nature* 217.5129 (1968), pp. 624–6.
- [193] R. Nielsen. "Molecular signatures of natural selection". *Annu Rev Genet* 39 (2005), pp. 197–218.
- [194] E. Pauws, A. H. van Kampen, S. A. van de Graaf, J. J. de Vijlder, and C. Ris-Stalpers. "Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis". *Nucleic Acids Res* 29.8 (2001), pp. 1690–4.
- [195] X. Wu, M. Liu, B. Downie, C. Liang, G. Ji, et al. "Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation". *Proc Natl Acad Sci U S A* 108.30 (2011), pp. 12533–8.
- [196] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, et al. "The transcriptional landscape of the yeast genome defined by RNA sequencing". *Science* 320.5881 (2008), pp. 1344–9.
- [197] E. Beaulieu, S. Freier, J. R. Wyatt, J. M. Claverie, and D. Gautheret. "Patterns of variant polyadenylation signal usage in human genes". *Genome Res* 10.7 (2000), pp. 1001–10.
- [198] C. Trapnell, L. Pachter, and S. L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq". *Bioinformatics* 25.9 (2009), pp. 1105–11.
- [199] M. D. Katinka, S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, et al. "Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*". *Nature* 414.6862 (2001), pp. 450–3.

- 
- [200] S. Douglas, S. Zauner, M. Fraunholz, M. Beaton, S. Penny, et al. "The highly reduced genome of an enslaved algal nucleus". *Nature* 410.6832 (2001), pp. 1091–6.
- [201] C. H. Slamovits and P. J. Keeling. "A high density of ancient spliceosomal introns in oxymonad excavates". *BMC Evol Biol* 6 (2006), p. 34.
- [202] C. W. Williams and H. G. Elmdorf. "Identification and analysis of the RNA degrading complexes and machinery of *Giardia lamblia* using an in silico approach". *BMC Genomics* 12 (2011), p. 586.
- [203] X. S. Chen, D. Penny, and L. J. Collins. "Characterization of RNase MRP RNA and novel snoRNAs from *Giardia intestinalis* and *Trichomonas vaginalis*". *BMC Genomics* 12 (2011), p. 550.
- [204] J. M. Feng, J. Sun, D. D. Xin, and J. F. Wen. "Comparative Analysis of the 5S rRNA and Its Associated Proteins Reveals Unique Primitive Rather Than Parasitic Features in *Giardia lamblia*". *PLoS One* 7.6 (2012), e36878.
- [205] F. Xu, J. Jerlstrom-Hultqvist, and J. O. Andersson. "Genome-Wide Analyses of Recombination Suggest That *Giardia intestinalis* Assemblages Represent Different Species". *Mol Biol Evol* (2012).
- [206] L. Clarke, X. Zheng-Bradley, R. Smith, E. Kulesha, C. Xiao, et al. "The 1000 Genomes Project: data management and community access". *Nat Methods* 9.5 (2012), pp. 459–462.
- [207] K. Sandahl, S. Botero-Kleven, and U. Hellgren. "[Chagas' disease in Sweden—great need of guidelines for testing. Probably hundreds of seropositive cases, only a few known]". *Lakartidningen* 108.46 (2011), pp. 2368–71.
- [208] M. Aslett, C. Aurecochea, M. Berriman, J. Brestelli, B. P. Brunk, et al. "TriTrypDB: a functional genomic resource for the Trypanosomatidae". *Nucleic Acids Res* 38.Database issue (2010), pp. D457–62.
- [209] I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, et al. "Tablet—next generation sequence assembly visualization". *Bioinformatics* 26.3 (2010), pp. 401–2.
- [210] F. Raymond, S. Boisvert, G. Roy, J. F. Ritt, D. Legare, et al. "Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species". *Nucleic Acids Res* 40.3 (2012), pp. 1131–47.
- [211] T. Downing, H. Imamura, S. Decuypere, T. G. Clark, G. H. Coombs, et al. "Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance". *Genome Res* 21.12 (2011), pp. 2143–56.
- [212] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels". *Bioinformatics* 28.8 (2012), pp. 1086–92.
- [213] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome". *Nat Biotechnol* 29.7 (2011), pp. 644–52.

- 
- [214] A. Regoes, D. Zourmpanou, G. Leon-Avila, M. van der Giezen, J. Tovar, and A. B. Hehl. "Protein import, replication, and inheritance of a vestigial mitochondrion". *J Biol Chem* 280.34 (2005), pp. 30557–63.
- [215] R. Neringer, Y. Andersson, and R. Eitrem. "A water-borne outbreak of giardiasis in Sweden". *Scand J Infect Dis* 19.1 (1987), pp. 85–90.
- [216] J. Muller, S. Ley, I. Felger, A. Hemphill, and N. Muller. "Identification of differentially expressed genes in a *Giardia lamblia* WB C6 clone resistant to nitazoxanide and metronidazole". *J Antimicrob Chemother* 62.1 (2008), pp. 72–82.
- [217] E. Ringqvist, L. Avesson, F. Soderbom, and S. G. Svard. "Transcriptional changes in *Giardia* during host-parasite interactions". *Int J Parasitol* 41.3-4 (2011), pp. 277–85.
- [218] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. "Computational methods for transcriptome annotation and quantification using RNA-seq". *Nat Methods* 8.6 (2011), pp. 469–77.
- [219] J. Ankarklev, S. G. Svard, and M. Lebbad. "Allelic sequence heterozygosity in single *Giardia* parasites". *BMC Microbiol* 12.1 (2012), p. 65.
- [220] Z. Faghiri and G. Widmer. "A comparison of the *Giardia lamblia* trophozoite and cyst transcriptome using microarrays". *BMC Microbiol* 11 (2011), p. 91.
- [221] J. O. Andersson. "Double peaks reveal rare diplomonad sex". *Trends Parasitol* 28.2 (2012), pp. 46–52.
- [222] R. S. Gupta and G. B. Golding. "The origin of the eukaryotic cell". *Trends Biochem Sci* 21.5 (1996), pp. 166–71.

